

HIGH SCHOOL MATHEMATICS TEACHERS' EVOLVING
UNDERSTANDING OF *COMPARING DISTRIBUTIONS*

by

Sandra R. Madden

A Dissertation
Submitted to the
Faculty of the Graduate College
in partial fulfillment of the
requirements for the
Degree of Doctor of Philosophy
Department of Mathematics
Advisor: Christian R. Hirsch, Ph.D.

Western Michigan University
Kalamazoo, Michigan
December 2008

Copyright by
Sandra R. Madden
2008

ACKNOWLEDGMENTS

I am honored and humbled to be writing these pages. I have been personally blessed with family, friends, mentors, and colleagues without whom I would not likely have conducted this research nor constructed this document. As I have reflected on the many people who have given so generously their time as well as intellectual, spiritual, and emotional support, I am again reminded of the blessings bestowed upon me. There are many people to whom I am eternally grateful.

First to my family, I could not have done this without your love and support. To my husband, Scott, thank you for supporting me and being, at times, a single-parent to our four children. Thanks for putting up with my grumpiness and for going above and beyond the call of duty. I know I wasn't always the most demonstrably appreciative of your contributions and I hope you know that I know that I could not have done this without your tremendous help. To my oldest daughter Amanda, thanks for being my rock and for not letting me off easily—you go girl. To my youngest daughter Chelsea, thanks for making me stop and take notice of things I may have been neglecting, for pointing them out to me in not-so-subtle ways and for keeping life interesting. To my oldest son Keegan, thank you for checking on my progress everyday and helping to make sure I finished this—you are mature beyond your years. To my youngest son Logan, thank you for being so flexible and adorable that you could deal with your part-time mom and still make her feel spectacular. To my mom and dad, thanks for helping me to become an independent thinker with a spirited outlook on life. Thanks for sharing all of my many

Acknowledgments—Continued

trials and tribulations and for being there to celebrate as well as pick up the pieces! To my extended family, I thank you all for your patience and support over the years. To my whole crew, I love you all dearly and cannot possibly thank you enough.

Second, there are a number of people who have become my friends and colleagues over the course of the Ph.D.-seeking adventure. To Robert Laing, thank you for encouraging and believing in me, for your contributions to this work, and for your friendship over the years. To Lisa Kasmer, thank you for your friendship and willingness to commiserate about the toils of dissertating—we should have invested in Alltel. To Rose Martin for years of professional collaboration and work on the “Roadshow,” thanks for your friendship, your company over literally thousands of miles, your perspective and encouragement. To Cynthia Halderson, my angelic friend who kept me going when I wanted to quit—you were an inspiration to me and I cannot thank you enough. To John Tanis, my professor of physics friend and confidante, thanks for long talks and lunches, for encouraging me, and for making life on the first floor occasionally palatable. To Hope Smith, thanks for helping me in so many ways over the years, I truly appreciate all of your words of wisdom, kindness, and incredible skills.

Third, to all of the teachers with whom I have been honored to interact as professional development facilitator, researcher, mentor, coach, and colleague, thank you for allowing me to share parts of your lives—we impacted each other in profound ways, didn’t we? Your commitment to teaching and making the world a better place provided a space for me to challenge theories of mine and others, test and refine my practice, and realize how amazing and powerful relationships are in the process of learning.

Acknowledgments—Continued

Fourth, to my dissertation committee, thank you. Thank you for your guidance, your patience, and your willingness to let me try to fly—including the occasional crash-and-burn. Christian Hirsch, you have been the mentor-of-mentors for me. You have inspired me from the time I was an undergraduate student of yours in Math 340 (23 years ago) and I have the highest regard for your work and your standards. Thank you for helping to shape how I have come to think about mathematics teaching and learning. Thank you for your fierce dedication to excellence. Thank you for your relentless pursuit of positive change in mathematics education, even amid strong critics. Thank you for encouraging my Fellowship in the Center for the Study of Mathematics Curriculum and for “suggesting” I work on the National Standards Analysis Project, among other things. Thanks for allowing me to pursue statistics education research. Mostly, thanks for believing in me. Steven Ziebarth, thank you for your time, patience, and willingness to read and react to so much of my work over the years. Thanks for letting me co-teach Math 654 with you and for supporting statistics education research. Thanks for letting me vent, for being my sounding board, and for providing important feedback at critical times that kept me going. Jon Davis, thank you for being willing to be part of my committee and for agreeing to let me attempt this project even though you knew it was too big. Magdalena Niewadomska-Bugaj, thanks for being my teacher of statistics and for being willing to venture into the realm of statistics education with me. To all of you, I am so very appreciative of your guidance and support. Thank you.

Finally, I would like to thank a special person who encouraged girls to be smart, independent, and empowered. To my gymnastics coach for many years, mother of five,

Acknowledgments—Continued

woman with many hats, and extraordinary person in my life, Alice Buchalter, thank you. Thank you for making me eat crepes at the Magic Pan restaurant, for making me dance every Tuesday, and for accepting no less than my best. Thank you for making me part of your family and for your amazing impact on my life. Thank you for expanding my world and helping me to envision possibilities beyond those I may have imagined.

To all of you and others, you have profoundly and positively impacted my life. This is for you.

Sandra R. Madden

TABLE OF CONTENTS

ACKNOWLEDGMENTS	ii
LIST OF TABLES	xii
LIST OF FIGURES	xvii
CHAPTER	
I. INTRODUCTION	1
A Call for Statistical Literacy	1
Statistical Literacy, Thinking, and Reasoning Defined	2
“Big Ideas” in Statistics	3
Statistics in High School	4
The State of Teacher Professional Development in Statistics	7
Technology for Teaching and Learning Statistics	9
The Role of Design Research	10
The Present Study	11
Statement of the Research Questions	12
Overall Description of the Study	12
Participants	12
Data Collection	13
Analysis of Data	15
Significance of the Study	16

Table of Contents—Continued

CHAPTER

II. LITERATURE REVIEW	17
Statistics Education Research: General Background	18
<i>Comparing Distributions</i> : General Background	20
<i>Comparing Distributions</i> : Student and Teacher Understanding	23
Students’ and Teachers’ Understanding of Distributions	32
Students’ and Teachers’ Understanding of Variability	34
Students’ and Teachers’ Understanding of Sampling Distributions	39
Informal Inference: What Is It? and What Do Students Know?	46
Technology to Support Statistics Learning	49
Use of Simulations	51
Use of Resampling Methods	54
Influence of Prior Research on the Present Study	56
Design Considerations for Supporting High School Mathematics Teachers’ Statistical Learning through Professional Development	57
The Importance of Professional Development for Supporting Teachers’ Understanding of Statistical Concepts	58
Theoretical Perspectives	59
Design Research as an Appropriate Methodology in Statistics Professional Development Research	63
Toward a Basis for Statistical Content Inclusion in Professional Development for High School Mathematics Teachers	66

Table of Contents—Continued

CHAPTER

III. METHODOLOGY	69
Theoretical Framework	69
Research Sites	73
The Setting	75
Participants	77
Interview Participants	80
Data Collection/Instrumentation	82
Analysis of Data	84
Answering the Research Questions	86
Organization of Remaining Chapters	92
IV. DESIGN AND IMPLEMENTATION OF THE PROFESSIONAL DEVELOPMENT INTERVENTION	93
What Is a Hypothetical Learning Trajectory?	93
Rationale for the Hypothetical Learning Trajectory	94
Summary of Pre-Assessment Responses	94
Summary of Initial Interview Responses	99
Designing the Hypothetical Learning Trajectory	101
Relationship Between the HLT and Assessment	109
The Emergent Learning Trajectory	110
Select Episodes from the Enacted Learning Trajectory	117
Summary of Themes Emerging from Retrospective Analysis of Professional Development Programs	167

Table of Contents—Continued

CHAPTER

Coordinating Content, Technology, and Professional Development Intervention Characteristics	168
V. RESULTS	172
Content Assessment Results	172
Establishing Comparability Across Sites	174
Aggregate Results (All 10 Items)	175
Aggregate Results (Eight Parallel Items)	180
Aggregate Results: Additional Comparisons	183
Item-Specific Analyses	186
Summary of Item-Specific Analyses	237
Summary of Content Assessment Results	239
Interview Results	241
Descriptions of Interview Teachers and Pre-Interview Summaries	242
Post-Interviews: Summary and Emerging Themes	267
Written Reflection Results	285
Reflection 1	286
Reflection 2	288
Reflection 3	292
Reflection 4	301
Reflection 5	302
Summary of Reflections	320

Table of Contents—Continued

CHAPTER

Transitioning to Chapter VI	321
VI. CONCLUSION	322
Answering the Research Questions: Bracketing the Researcher’s View	323
Question 1: Knowledge of <i>Comparing Distributions</i>	325
Question 2: Resampling Techniques and Dynamic Statistical Tool Usage	334
Question 3: Characteristics of Professional Development	347
Discussion	351
Dynamic Technology Scaffolding: A Design Principle for Supporting Teachers’ Understanding of <i>Comparing Distributions</i>	353
Dynamic Technology Scaffolding: What Is It and Why Should We Care?	363
Discussion Summary	365
Limitations and Implications	366
Suggestions for Future Research	368
Concluding Thoughts	370
REFERENCES	371
APPENDICES	
A. Pre-Assessment	384
B. Post-Assessment	395
C. Pre-Intervention Interview Protocol	407

Table of Contents—Continued

APPENDICES

D.	Post-Intervention Interview Protocol	410
E.	Written Reflection Prompts	415
F.	Assessment Scoring Rubrics	421
G.	Professional Development Facilitator Guide Pages	432
H.	Mapping Professional Development Activities to Assessments.....	449
I.	Human Subjects Institutional Review Board Approval	454

LIST OF TABLES

1. Taxonomy for Classifying Levels of Reasoning When Comparing Two Groups	30
2. Framework for Levels of Understanding of Variability	38
3. Stages of Development in Students' Statistical Reasoning about Sampling Distributions	44
4. Big Ideas, Sub-Ideas, and Considerations Related to "Comparing Distributions"	67
5. Coordinating Multiple Frameworks—Comparing Distributions, Variability, and Sampling Distributions	70
6. Regional Site Distribution and 2006-2007 Demographic Data for Schools of Participating Teachers	74
7. Data Collection Schedule for Summer 2006	76
8. Participating Teachers' Reported Certification	77
9. Teachers' Self-Reported Comfort Level for Big Ideas from Statistics	79
10. Background Characteristics of Interview Subjects	83
11. An Interpretive Framework for Analyzing Classroom Mathematical Activity and Learning	86
12. Explication and Chronology of the Initial Hypothetical Learning Trajectory	95
13. Number and Nature of Learning Activities with Potential to Support the Development of Big Ideas	102
14. Learning Activities as Sequenced and Enacted in PD Session 1	113
15. Learning Activities as Sequenced and Enacted in PD Session 2	114
16. Learning Activities as Sequenced and Enacted in PD Session 3	115

List of Tables—Continued

17.	Summary of Factors Potentially Impacting Teachers’ Understanding of Comparing Distributions	169
18.	Summary of the Enacted Evolution of Content, Technology Use, and Characteristics of Professional Development	170
19.	Correspondence Between Assessment Tasks and Item Clusters	173
20.	Aggregate Pre-Assessment Scores Across Sites	175
21.	Distribution of Pre-Assessment Scores by Item	176
22.	Aggregate Post-Assessment Scores Across Sites(10 Items)	177
23.	Distribution of Post-Assessment Scores by Item	178
24.	Pre- to Post-Assessment Gain-Scores (10 Items)	179
25.	Content Assessment Mean Pre-, Post-, and Gain-Scores by Professional Development Site	185
26.	Item 1: Sample Responses and Scores	189
27.	Item 2: Sample Responses and Scores	195
28.	Item 3: Frequency of Responses on Pre- and Post-Assessment	198
29.	Item 3: Change in Responses Pre- to Post-Assessment	198
30.	Incorrect Answers and Explanations for Post-Assessment Item 3	198
31.	Item 4: Response Patterns Pre- to Post-Assessment	202
32.	Comparison Between Monte Carlo Simulation and Randomization Testing Procedures	206
33.	Item 6: Sample Pre-Assessment Responses	207
34.	Item 6: Comparison of Performance on Post-Assessment Item 6 with Number of Statistics Course Taken	209
35.	Item 7: Sample Responses and Scores	212

List of Tables—Continued

36.	Item 8: Question by Question Comparison Pre- to Post-Assessment	218
37.	Item 8: Sample Responses and Scores	219
38.	Item 8: Pre- and Post-Assessment Comparisons for Two Teachers	221
39.	Item 9: Pre- and Post-Assessment Comparisons for Four Teachers	226
40.	Item 10: Rubric for Scoring	229
41.	Item 10: Pre- and Post-Assessment Results from Tasks 14 through 20	231
42.	Item 10: Total Number of Tasks Correct by Teacher	232
43.	Item 10: Pre- and Post Responses by Task (with correct answers in bold)	232
44.	Pre-, Post-, and Gain-Score Comparisons by Big Idea Cluster	239
45.	Teachers' Self-Reported Comfort Level with Statistical Ideas at Pre- and Post-Assessment	240
46.	Characteristics of Interview Teachers	241
47.	Interview Teachers' Pre-Assessment Scores by Item	268
48.	Interview Teachers' Post-Assessment Scores by Item	268
49.	Interviewed Teachers, Non-Interviewed Teachers, and All 56 Teachers: Comparison from Pre- to Post- and Gain-Scores	269
50.	Teachers' Self-Characterizations of Their Use of Technology	280
51.	Summary of Teachers' Interview Times and Change in Understanding on Post-Assessment by Sub-Idea and Overall	283
52.	Summary of Interview Teachers' Content Knowledge and Self-Reported Comfort Level with Statistics Pre- and Post-Intervention	284
53.	Reflection 1: Nature of Teachers' Responses by Site	287
54.	Reflection 1: Frequency of Big Ideas for which Teachers Reported Improved Understanding	287

List of Tables—Continued

55.	Reflection 2: Categorization of Responses for Parts A, B, C, and D	289
56.	Reflection 5: All Responses to Part D	291
57.	Reflection 3A: Nature of Teachers' Responses by Site	293
58.	Reflection 3A: Frequency of Teacher-Reported Improved Understanding of Measures of Center	294
59.	Reflection 3B: Nature of Teachers' Responses by Site	295
60.	Reflection 3B: Frequency of Teacher-Reported Improved Understanding of Variability	296
61.	Reflection 3C: Nature of Teachers' Responses by Site	297
62.	Reflection 3C: Frequency of Teacher-Reported Improved Understanding of Bias	297
63.	Reflection 3D: Nature of Teachers' Responses by Site	298
64.	Reflection 3D: Frequency of Teacher-Reported Improved Understanding of Design of Experiments	299
65.	Reflection 3E: Frequency of Teacher-Reported Questions of Understanding	300
66.	Reflection 4B: Frequency and Samples of Responses by Level	303
67.	R4-C and R5-A: Sample Teacher Responses to Reflection Questions about "Significant Differences"	305
68.	Teachers Self-Reported Changes or Growth in Ideas about Comparing Distributions with Pre- and Post-Beliefs and Assessment Results	309
69.	Reflection 5: Categorization of Responses for Part C	313
70.	Teachers' Self-Reported Ratings for Their Understanding of the Randomization Test Across Reflection Opportunities	316
71.	Change in Teachers' Self-Reported Rating for Understanding of the Randomization Test Across Reflection Opportunities	317

List of Tables—Continued

72.	Teachers' Self-Reported Ratings for Their Feelings about Their Learning of <i>Fathom2</i>	320
73.	Resampling Techniques and Dynamic Statistical Technology: Contributing to Teachers' Understanding of Comparing Distributions	338
74.	The Randomization Test Across Multiple Technological Environments	354

LIST OF FIGURES

1.	Synergy in statistical education	19
2.	A conjectured conceptual relationship between big statistical ideas	68
3.	Distribution of participating teachers' number of years teaching mathematics	77
4.	Teachers' average comfort level with statistical ideas compared to experience, low = 1, high = 5	80
5.	Mapping content to assessment	110
6.	Graphical representations from the handspan investigation	133
7.	Sliders in <i>Fathom2</i> to dynamically model parameter changes to the standard normal density function	154
8.	Sample activity: Is there a relationship between the sample size and the sampling distribution's standard deviation?	158
9.	Teachers' mathematical models from Activity 3.4	161
10.	The randomization test for the Physician's Health Study	164
11.	Comparison of aggregate pre-assessment scores across sites with ANOVA results	175
12.	Comparison of aggregate post-assessment scores across sites with ANOVA results	178
13.	Comparison of aggregate gain-scores across sites with ANOVA results	180
14.	Comparison of pre-, post-, and gain-score distributions and matched pairs <i>t</i> -test results (10 items)	180
15.	Comparison of pre-, post-, and gain-score distributions and matched pairs <i>t</i> -test results (8 parallel items)	181
16.	Gain-scores for 10 items compared to gain-scores for 8 parallel items	182

List of Figures—Continued

17.	Item-by-item post-assessment means against pre-assessment means	182
18.	Comparisons of pre-, post-, and gain score distributions by number of statistics courses taken	184
19.	Item 1: Pre-, post-, and gain-score comparisons	188
20.	Item 2: Pre-, post-, and gain-score comparisons	194
21.	Item 3: Pre- and post-assessment task prompts	196
22.	Item 3: Pre-, post-, and gain-score comparisons	197
23.	Item 4: Pre-, post-, and gain score comparisons	200
24.	Item 5: Pre-, post-, and gain score comparisons	205
25.	Item 6: Pre-, post-, and gain-score comparisons	207
26.	Item 7: Pre-, post-, and gain-score comparisons	211
27.	Item 7: Summary statistics and side-by-side boxplots from the task.....	214
28.	Item 8: Pre-, post-, and gain-score comparisons	217
29.	Item 9: Pre-, post-, and gain-score comparisons	224
30.	Item 10: Pre-, post-, and gain-score comparisons.....	230
31.	Item 10: Comparison between number correct on post-assessment (horizontal axis) to pre-assessment score (vertical axis)	237
32.	Item 10: Post-assessment scores disaggregated by number of statistics courses taken	238
33.	R2-E: Distribution of teachers' ratings of the ease of use of <i>CPMP-Tools</i>	292
34.	R5-A and R4-C: Summary of reflection scores and levels for “significant difference”	307
35.	R5-F: Distribution of teachers' ratings of understanding the relationship between sample size and sampling distributions	314

List of Figures—Continued

36.	R5-F: Post-assessment score for Item 10 (Questions 14–20) compared to teachers’ self-assessed understanding of sampling distributions from Reflection 5-F	315
37.	Comparison of pre- and post-assessment distributions overall and by big idea	327
38.	A model for the relationships among content, representations, and technological tools supporting teachers’ understanding of comparing distributions	342
39.	A model of teachers’ evolving understanding of <i>comparing distributions</i>	351
40.	An example of the randomization distribution tool in <i>CPMP-Tools</i>	357
41.	An example of the randomization distribution mechanism in <i>Fathom2</i>	360
42.	Model of dynamic technology scaffolding	363

CHAPTER I

INTRODUCTION

A Call for Statistical Literacy

With the rapidly increasing power and capability of technology, the amount of data being generated and accumulated is increasing at an increasing rate. Data are frequently used to drive decision-making at the level of nations, states, local communities, businesses, schools, and individuals, and the need for statistically literate citizens has never been greater. According to Konold and Higgins (2003),

At the practical level, knowledge of statistics is a fundamental tool in many careers, and without an understanding of how samples are taken and how data are analyzed and communicated, one cannot effectively participate in most of today's important political debates about the environment, health care, quality of education, and equity. For those who have traditionally been left out of the political process, probably no skill is more important to acquire in the battle for equity than statistical literacy. (p. 193)

As early as 1923, it was suggested that statistics have a place in the secondary school curriculum (National Committee on Mathematical Requirements [NCMR], 1923). More recently, as evidenced through the progression of documents such as *A Nation At Risk* (Gardner et al., 1983); *The Mathematical Sciences Curriculum K-12: What Is Still Fundamental and What Is Not?* (Conference Board of the Mathematical Sciences [CBMS], 1982); NCTM's *Curriculum and Evaluation Standards for School Mathematics* (National Council of Teachers of Mathematics [NCTM], 1989) and *Principles and Standards for School Mathematics* (PSSM) (NCTM, 2000); *Before It's Too Late*

(National Commission on Mathematics and Science Teaching, 2000); *Ready or Not: Creating a High School Diploma that Counts* (American Diploma Project, 2004); and *Guidelines for Assessment and Instruction in Statistics Education* (GAISE) (Franklin et al., 2007), an increasing emphasis for K-12 education in the areas of data analysis, statistics, and probability has been strongly recommended. The *College Board Standards for College Success* (College Board, 2006b) and *Secondary Mathematics Benchmarks Progressions: Grades 7-12* (Achieve, Inc., 2007), both national documents with influence on college entrance examinations and state standards for high school mathematics, reflect the position that statistical preparation for students is no longer optional. In light of these recommendations, national- and state-level assessments have begun to reflect a statistical content focus (e.g., NAEP, SAT, ACT, MEAP).

Statistical Literacy, Thinking, and Reasoning Defined

The statistics education research field is still in its infancy and, as such, is still wrestling with concise definitions of terms such as *statistical literacy*, *statistical thinking*, and *statistical reasoning*. Definitions have been proposed and can be found in *Statistical Literacy, Thinking, and Reasoning: Goals, Definitions, and Challenges* (Ben-Zvi & Garfield, 2004b). Briefly,

statistical literacy includes the skills that might be used to understand statistical information or research results. *Statistical reasoning* is the way in which people reason with statistical ideas and make sense out of statistical information. *Statistical thinking* involves an understanding of why and how statistical investigations are conducted and the “big ideas” that underlie statistical investigations. (p. 7)

By encouraging statistical thinking and reasoning, statistical literacy is encouraged.

“Big Ideas” in Statistics

According to Moore (1990), the core elements of statistical thinking are:

1. The omnipresence of *variation* in processes. Individuals are variable; repeated measurements on the same individual are variable. The domain of a strict determinism in nature and in human affairs is quite circumscribed.
2. The need for *data* about processes. Statistics is steadfastly empirical rather than speculative. Looking at the data has first priority.
3. The design of *data production* with variation in mind. Aware of sources of controlled and uncontrolled variation, we avoid self-selected samples and insist on comparison in experimental studies. And we introduce planned variation into data production by use of randomization.
4. The *quantification* of variation. Random variation is described mathematically by *probability*.
5. The *explanation* of variation. Statistical analysis seeks the systematic effects behind the random variability of individuals and measurements. (p. 135)

According to Shaughnessy (2007), prior to 1999, there was little research on students’ conceptions of variability.

Since 1999, variability and related statistical ideas have emerged as centerpieces for research (Ben-Zvi & Garfield, 2004a). As evidenced by the recent and upcoming conference foci for the International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL), one can begin to sense the progression of the statistics education research community from grappling with definitions to investigating student reasoning about big ideas in the field. The following represent foci for biannual conference meetings for SRTL: Statistical Reasoning, Thinking, and Literacy (1999); Statistical Reasoning: Reasoning about data and distributions, variability, sampling, comparing distributions, and sampling distributions (2001); Reasoning about Variability (2003); Reasoning about Distribution (2005); Reasoning about Statistical Inference (2007); and The Role of Context and Evidence in Informal Inferential Reasoning (2009). This list provides a context for much of the literature reviewed in Chapter II.

Much recent research has been conducted on students' understanding of specific statistics concepts, but a small number of studies investigating teachers' understanding of statistical concepts have also been conducted (e.g., Makar & Confrey, 2004; Mickelson & Heaton, 2004). Until recently, statistics education research focused largely on that of college students' understanding, related largely to probability (Shaughnessy, 2007).

As will be explicated in Chapter II, there is a great deal of overlap among research studies of student reasoning about distributions, variability, sampling distributions, and informal inference. These concepts are intertwined through their embodiments in data analysis and inference; therefore, studying reasoning in these areas as isolated topics is problematic. *Comparing distributions* has emerged as a fruitful venue for engaging learners across age-bands in thoughtful pursuit of understanding distributions, variability, and sampling distributions, potentially leading to a better understanding of ideas related to inference (e.g., Bakker & Frederickson, 2005; Ben-Zvi, 2004; delMas, Garfield, & Chance, 1999; Konold, Pollatsek, Well, & Gagnon, 1997; Lehrer & Schauble, 2004; Makar & Confrey, 2004; Pfannkuch, Budgett, Parsonage, & Horring, 2004; Watson & Moritz, 1999).

Statistics in High School

The traditional sequence of mathematics coursework at the high school level in the United States has been Algebra I, Geometry, Algebra II, and Pre-calculus. For mathematically “able” students, calculus was the goal, with substantially less rigorous expectations set for students seen as “less able” (Jones & Coxford, 1970). Statistics has been incorporated into some of these courses through the use of data to drive

mathematical modeling and use of technology (Usiskin, 2003). Aside from the use of regression techniques for modeling, one sees very little attention to statistics in the traditional high school curriculum. Until recently, because curricula have historically been dominated by preparation for calculus, little room for statistics in high school has been available for most students (CBMS, 2001).

With respect to statistics in school, *PSSM* (NCTM, 2000) recommends that students in grades K-12 should

1. Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them.
2. Select and use appropriate statistical methods to analyze data.
3. Develop and evaluate inferences and predictions that are based on data.
4. Understand and apply basic concepts of probability. (p. 324)

In particular, *PSSM* recommends that grades 9-12 high school students should

gain a deep understanding of the issues entailed in drawing conclusions in light of variability. They will learn more sophisticated ways to collect and analyze data and draw conclusions from data in order to answer questions or make informed decisions in workplace and everyday situations. They should learn to ask questions that will help them evaluate the quality of surveys, observational studies, and controlled experiments. They can use their expanding repertoire of algebraic functions, especially linear functions to model and analyze data, with increasing understanding of what it means for a model to fit data well. In addition, students should begin to understand and use correlation in conjunction with residuals and visual displays to analyze associations between two variables. They should become knowledgeable, analytical, thoughtful consumers of the information and data generated by others. (NCTM, 2000, p. 325)

These recommendations are further refined and elaborated in the *Guidelines for Assessment and Instruction in Statistics* (GAISE), in which learner developmental levels A, B, and C are crossed with the statistical problem-solving processes of formulating questions, collecting data, analyzing data, and interpreting data (Franklin et al., 2007). The GAISE framework attends to the nature of variability and suggests focus on variability in increasingly sophisticated ways as level increases from A to B to C. The

authors of GAISE were cautious to state that though the developmental levels in the framework

may parallel grade levels, they are based on development in statistical literacy, not age. Thus a middle-school student who has had no prior experience with statistics will need to begin with Level A concepts and activities before moving to Level B. (p. 13)

Since Level C is loosely associated with high school level, high school teachers should understand relevant statistics leading up to and beyond Level C in order to support student learning of statistics.

In line with professional recommendations like those of NCTM, K-12 mathematics curricula developed with funding from the National Science Foundation (NSF) have incorporated data analysis and statistics as an integrated mathematical strand in the creation of their materials. For example, the Connected Mathematics Project, a grades 6-8 mathematics curriculum, contains units specifically devoted to support students' statistical reasoning (e.g., *Samples and Populations*; *Data Around Us*, *Data About Us*) (Lappan, Fey, Fitzgerald, Friel, & Phillips, 1998). Similarly, a high school NSF-funded curriculum project, the Core-Plus Mathematics Project (Coxford et al., 2003) first edition included units titled *Patterns in Data*, *Simulation Models*, *Patterns of Association*, *Patterns of Chance*, *Modeling Public Opinion*, *Patterns in Variation*, and *Binomial Distributions and Statistical Inference*. The second edition of Core-Plus Mathematics (Hirsch, Fey, Hart, Schoen, & Watkins, 2008) has incorporated an even stronger emphasis on developing statistical literacy for students and has developed a companion suite of technological tools to support student investigation of statistical concepts in the curriculum. The point is that the mathematics education field has recognized the necessity of instructional materials to support student understanding of

statistics. In particular, the field has finally acknowledged the importance of variability as the essence of why we study statistics at all (Moore, 1997). It will require a statistically literate teaching force to effectively implement these and other statistically-oriented instructional materials in classrooms.

Another fairly recent development includes the increasingly popular Advanced Placement (AP) Statistics course in high school. The number of annual AP Statistics exams taken in the United States continues to increase and has grown by more than a factor of 10 since its inception in 1997 (College Board, 2006a). The rate of growth seen in the annual number of AP Statistics examinations is more than double that of the number of AP Calculus AB and BC exams combined. With new choices for instructional materials and coursework, teachers' preparation and professional development opportunities should support teachers' developing conceptions of statistics.

The State of Teacher Professional Development in Statistics

Recognizing the Conference Board of Mathematical Sciences (2001) report, *The Mathematical Education of Teachers*, the importance of statistics in the K-12 curriculum, recommended that high school teachers' professional preparation include experience formulating questions, devising data collection protocols, and analyzing real data sets that result from their own investigations or from the data collection of others. It is recommended that high school teachers' preparation support the appreciation and understanding of the major themes of statistical practice:

- Exploring data: using a variety of standard techniques for organizing and displaying data in order to detect patterns and departures from patterns.
- Planning a study: using surveys to estimate population characteristics and designing experiments to test conjectured relationships among variables.

- Anticipating patterns: using theory and simulations to study probability distributions and apply them as models of real phenomena.
- Statistical inference: using probability models to draw conclusions from data and measure the uncertainty of those conclusions. (CBMS, 2001)

These recommendations have obvious statistical content-specific ramifications for both pre-service and in-service teacher programs.

Statistics, however, is a relatively new subject for many teachers, who have not had an opportunity to develop sound knowledge of the principles and concepts underlying the practices of data analysis that they now are called upon to teach. These teachers do not clearly understand the difference between statistics and mathematics. They do not see the statistics curriculum for grades pre-K-12 as a cohesive and coherent curriculum strand. These teachers may not see how the overall statistics curriculum provides a developmental sequence of learning experiences. (Franklin et al., 2007, p. 5)

Besides the issue of inadequate statistical content-knowledge, there is also an issue with pedagogical content knowledge (Shulman, 1986). In the United States,

Whether students are in rows working individually or sitting in groups, whether they have access to the latest technology or are working only with paper and pencil, they spend most of their time acquiring isolated skills through repeated practice. (Stigler & Hiebert, 1999, p. 11)

This communicates a vision that teachers in the United States widely believe mathematics is largely dominated by memorization of algorithms, transmission of ideas, and practicing procedures with students as passive recipients of knowledge. This vision stands in stark contrast to the recommendations of NCTM (1989, 2000), but is still prevalent in high school mathematics classrooms. According to Garfield and Ben-Zvi (2004), “more studies are needed that explore how to equip school teachers at all levels with appropriate content knowledge and pedagogical knowledge, and to determine what kind of guidance they need to successfully teach these [statistical] topics” (p. 404). Furthermore, according

to Shaughnessy (2007), “More research is needed on teachers’ conceptions of statistics” (p. 1000).

Technology for Teaching and Learning Statistics

In an environment of exploratory data analysis (EDA) and statistical investigation, the need for technology is great and the tools available for statistical investigation have evolved tremendously over the past decade. A number of technological tools are available for doing statistics, such as SAS, Minitab, and SPSS, but these tools, though powerful for statistical analysis, are less appropriate for learners embarking on studying statistics in the spirit of EDA. In particular, the development of *Fathom2* (Key Curriculum Press, 2005a) and *Tinkerplots* (Key Curriculum Press, 2005b) have revolutionized the ways in which students and teachers may interact with data. Both are dynamic statistical tools which allow users to flexibly move between representations of data and to explore connections between multiple, hot-linked representations (Zbiek, Heid, Blume, & Dick, 2007). Bakker (2002) described *Fathom2* and *Tinkerplots* as “landscape-type” tools because of the flexible and open ways in which users may interact with data and multiple representations. He contrasted landscape-type tools with “route-type tools,” which are more carefully structured and typically designed to support specific learning goals. Route-type tools may include Java applets (e.g., *CPMP-Tools*) and other specially-designed software environments (e.g., *Mini-tools*) constructed to support student learning.

The Role of Design Research

As a field, statistics education has moved beyond the study of student misconceptions of probability and seeks to better understand how to support the statistical development of teachers and students (Shaughnessy, 2007). Design research is particularly appropriate for the development and investigation of innovative approaches to teaching and learning (Brown, 1992). A number of researchers have found it productive for engineering and testing approaches to supporting the learning of various statistical concepts with learners representing a continuum of students and teachers (e.g., Abrahamson & Wilensky, 2007; Bakker, 2004; Brown, 1992; P. Cobb, Confrey, DiSessa, Lehrer, & Schauble, 2003; P. Cobb & McClain, 2004; P. Cobb, McClain, & Gravemeijer, 2003; Lehrer & Schauble, 2004; Makar, 2004). According to Cobb, Confrey, et al. (2003),

1. The purpose of design experimentation is to develop a class of theories about both the process of learning and the means that are designed to support that learning
2. Design experiments are typically test-beds for innovation
3. Design experiments are prospective and reflective as they create conditions for developing theories yet must place these theories in harm's way
4. Design is iterative as conjectures are tested and refuted, cycles of invention and revision dominate the process
5. Theories developed during the process of experiment are humble not merely in the sense that they are concerned with domain-specific learning processes, but also because they are accountable to the activity of design.

As will be further discussed in Chapter IV, the current study builds on past research on student and teacher understanding of *comparing distributions*. The use of simulations and resampling techniques in the context of statistical investigation by high school teachers while *comparing distributions* will be investigated. A design experiment is appropriate because it entails “both ‘engineering’ particular forms of learning and

systematically studying those forms of learning within the context defined by the means of supporting them” (P. Cobb, Confrey, et al., 2003, p. 9). Design experiments are structured, subject to testing and revision of theories, through successive iterations.

The Present Study

Of studies investigating teachers’ statistical conceptions, many have been in the context of elementary school or with particularly small groups of volunteers (Shaughnessy, 2007). The current study begins to fill a need in the literature by exploring the statistical conceptions of a fairly large group of high school mathematics teachers ($n = 56$) in the context of teacher professional development. The GAISE framework and other recent curriculum documents are useful for examining the nature of concepts appropriate for high school students in order to consider potential support for the statistical development of teachers.

Statistically worthwhile investigations frequently involve the use of technology. Because research on student statistical thinking and learning in new powerful software environments is just beginning to emerge, the current study contributes to what is known about how teachers interact with these types of technologies as they learn about *comparing distributions*. As will be described in Chapter IV, the professional development intervention was successively refined during three rounds of experimentation in reasonably well-defined circumstances. Through the careful study of teachers’ learning in environments such as those engineered in this study, this research may advance both knowledge and practice for the benefit of learners of statistics.

Statement of the Research Questions

This study focuses on whether and in what ways high school mathematics teachers' understanding of *comparing distributions* may be supported in the context of professional development. The goal of the professional development was to improve teachers' statistical thinking and reasoning and to support their developing facility with tools for learning and teaching statistics. The professional development intervention had the potential to help to transform teachers' ideas of learning statistics in an environment of challenging authentic statistical tasks, appropriate technological tools, and a spirit of collaboration.

The following research questions were investigated:

1. What do high school mathematics teachers know about *comparing distributions*?
2. How do professional development experiences with resampling techniques and dynamic statistical tools, as described in this study, shape what teachers know about *comparing distributions*?
3. What characteristics of professional development for high school mathematics teachers contribute to their understanding of *comparing distributions*?

Overall Description of the Study

Participants

Fifty-six (56) high school mathematics teachers from 23 school districts participating in a state-level Mathematics and Science Partnership (MSP) project were the participants for this study. The MSP was a two-year professional development and

research project designed to investigate ways to support middle and high school teachers' mathematical and pedagogical content knowledge and was the umbrella project under which this study was situated. The 56 teachers were distributed across three distinct regions within one state in the Midwest. Teachers were not randomly assigned to groups, but rather each group was determined by geography and teachers' abilities to attend professional development in a particular geographic location. Generally, teachers from a specific geographical area were clustered in a professional development site. One of the reasons for the MSP project was to provide professional development support to teachers in regions of the state which had been historically underserved due to distance considerations. The professional development that was the focus of this study was a four-day, technology-intensive, statistically-focused summer workshop for high school mathematics teachers. The four-day session was repeated, with minor modifications based upon the previous iteration, in each of three geographic locations.

Data Collection

A 21-item statistics pre-assessment was administered to all 56 teachers approximately two months prior to the professional development intervention. The instrument contained constructed-response, multiple-choice, and survey questions designed to assess the statistics background of teachers as well as some perceptions about their competence in specific areas of statistics. Teachers' responses to items were entered into an Excel spreadsheet for analysis. Pre-assessment results were used to select three teachers from each of the three professional development sites for pre-professional development interviews. The interviewees were selected to represent relatively high,

medium, and low levels of understanding of *comparing distributions*, based on the pre-assessment. These interviews were designed to provide the opportunity to probe teacher answers to the pre-assessment questions as well as to inform the design of the professional development intervention. The interviews were audio and videotaped and transcribed for analysis using NVivo software.

The four-day professional development program was conducted and videotaped at each of the three professional development sites. The researcher facilitated all professional development sessions while a second mathematics educator videotaped the sessions, took field notes, and co-reflected during and after each day with the researcher regarding the intervention.

At five distinct times during the four days of professional development, teachers responded to written reflection prompts regarding (1) their understanding of statistical ideas, and (2) learning issues they may be having. These written reflections were collected and coded by the researcher and entered into an Excel spreadsheet for analysis. The teachers were then post-assessed on the final day of the professional development program using an assessment instrument similar to that used at the time of the pre-assessment. Responses were entered into an Excel spreadsheet for analysis.

Post-professional development interviews were conducted with the same teachers who had been interviewed prior to the professional development to provide opportunity to probe their thinking on the assessment as well as to share their thinking on the professional development experience. These interviews were handled in the same way as the pre-intervention interviews.

Analysis of Data

The study utilized mixed methods. Pre- and post-assessment items were first scored using item-specific four-level scoring rubrics designed to represent levels of understanding of *comparing distributions*, variability, and sampling distributions (Chance, delMas, & Garfield, 2004; Makar & Confrey, 2004; Watson, Kelly, Callingham, & Shaughnessy, 2003). Quantitative analyses (descriptive statistics, ANOVA, paired *t*-tests, non-parametric procedures) were used to determine whether changes from pre- to post-assessment were significant. These analyses included a number of comparisons within and between professional development groups. Further qualitative analysis was used to investigate the nature of the change in responses from pre- to post-assessment.

Multiple qualitative research techniques were utilized in this study, but grounded theory (Strauss & Corbin, 1998) was used to generate, interrogate, and build theory. Numerous qualitative data coding and analysis procedures were required to categorize, sort, and reduce the data in search of patterns and explanations. Transcribed interviews were analyzed qualitatively using NVivo software for coding and theory building. Videotapes of the professional sessions were reviewed, chronologically catalogued, and analyzed in order to confirm or refute developing theories from the pre/post-assessments and interviews. Small segments of the videotape were transcribed to provide excerpts used to exemplify and clarify classroom activity and the development of sociomathematical norms. Written teacher reflections were analyzed qualitatively by prompt and across time to further contribute to theory building and to document teachers' evolving understanding of *comparing distributions*.

Significance of the Study

The design experiment conducted during this study was used to investigate a professional development intervention with potential for contributing to our understanding of how high school teachers come to think and reason in statistically powerful ways. In particular, the use of *comparing distributions* in a technology-rich environment was posited to have the potential to support multiple statistical connections, particularly among the closely related concepts of distribution, variability, and sampling distributions. The process began with a thought experiment to develop a hypothetical learning trajectory for professional development that led to creating an intervention with the goal of supporting teachers' developing conceptions of *comparing distributions*. It was hypothesized that through the use of resampling techniques and dynamic statistical software to model problem situations, teachers might view statistical ideas from a sense-making perspective. By carefully studying the enactment of the intervention in the professional development environment, a model was developed with potential for supporting teachers' statistical thinking and reasoning generally, as well as beginning to transform teachers' beliefs about mathematics and statistics learning toward a more sense-making and constructivist perspective.

CHAPTER II

LITERATURE REVIEW

As will be more fully detailed in Chapter III, this study may be categorized as design research in the context of professional development for high school mathematics teachers. Due to the nature of the study, literature related to (1) student and teacher understanding of statistical concepts, (2) the use of simulations and resampling using technology to learn statistics, (3) teacher professional development, and (4) design research is reviewed.

Because understanding *comparing distributions*, as a statistical big idea, is a focus of this study, literature related to teachers' and students' understanding of *comparing distributions* is reviewed. The review includes research on teachers' and students' understanding of the related concepts of distribution, variability, sampling distributions, and informal inference, and their relationship to *comparing distributions*. Because of the complex relationship among big ideas related to *comparing distributions* and research about them, sections and sub-sections in this chapter contain some overlap. The use of technology supporting learning of statistics through simulation and resampling techniques is also reviewed. Studies reviewed are limited to those most closely related to students' and teachers' understanding of statistical ideas in order to better frame the present study and to suggest direction for the design of the professional development intervention which is elaborated in Chapter IV. Further support for the design of the

intervention and the study, in general, comes from literature reviewed about teacher professional development and design research in statistics education related areas.

The chapter is organized in four major sections, three of which are further subdivided within each section: (1) Statistics Education Research: General Background; (2) *Comparing Distributions*: General Background; (3) Technology to Support Statistics Learning; and (4) Design Considerations for Supporting High School Mathematics Teachers' Statistical Learning. As the literature was reviewed, big ideas and sub-ideas were accumulated for use during the design of the intervention and to assist with the conceptualization of the study (see Table 4 near the end of the chapter). A conceptual framework (see Figure 2) that emerged from the collective review of the literature and provided structure for the construction of the hypothetical learning trajectory for this study is presented at the end of the chapter.

Statistics Education Research: General Background

Building on early work by Tukey (1977) with exploratory data analysis (EDA) techniques, recent changes in statistics education including increases in the availability, power, and use of technology in the practice of statistics; increased awareness of students' deficiencies with respect to thinking and reasoning statistically; and concerns with the inadequate preparation of teachers to teach statistics; have prompted efforts to change the teaching of statistics (Ben-Zvi & Garfield, 2004b). Moore (1997) suggested that "the most effective learning takes place when content (what we want students to learn), pedagogy (what we do to help them learn), and technology reinforce each other in

a balanced manner” (p. 124). He suggested a multi-directional triad between content, pedagogy, and technology as shown in Figure 1.

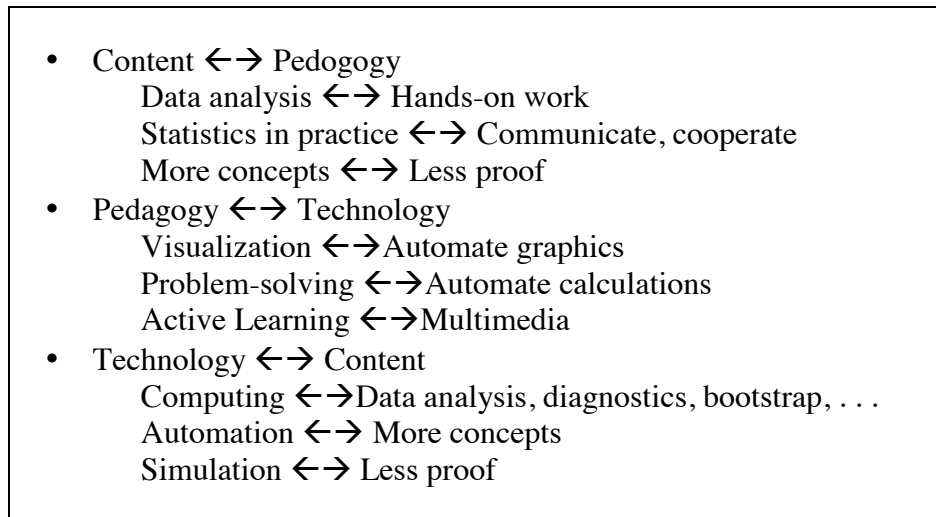


Figure 1. Synergy in statistical education.

(Note. From “New Pedagogy and New Content: The Case of Statistics,” by D. S. Moore, *International Statistical Review*, Vol. 65, No 2, p. 129.)

Moore realized the synergistic relationship between content, pedagogy and technology in teaching statistics. He strongly advocated the use of technology for learning statistics through simulations and resampling. At the time of Moore’s writing, software existed for doing statistics, but pedagogically-appropriate software for supporting the learning of statistics was still being developed (Bakker, 2002; P. Cobb, 1999; P. Cobb, McClain, et al., 2003; Konold, 1994). As will be described in further detail later in this chapter, contemporary tools for learning statistics include widely available Java-based software such as *CPMP-Tools* (Keller, 2006), stand-alone software packages such as *Tinkerplots* (Key Curriculum Press, 2005b), and *Fathom2* (Key Curriculum Press, 2005a). These tools have revolutionized the ways in which teachers and students may interact with data and statistical investigations due to their dynamic and flexible capabilities. Research is just beginning to tap into the potential of these learning

environments and accumulate evidence about the ways in which they may affect students' statistical thinking and reasoning (Shaughnessy, 2007).

Big ideas in statistics have surfaced as potential areas for organizing statistics instruction and for areas of research as they have in other areas of educational research (Garfield & Ben-Zvi, 2004). Garfield and Ben-Zvi (2004) suggest that statistical big ideas include data, distribution, trend, variability, models, association, samples and sampling, and inference, and recognize that other big ideas, for example, shape, may be productive to consider. Notably absent from this list, but potentially useful as a big idea in its own right as well as one with potential for supporting connections between multiple other big ideas, is that of *comparing distributions*. “We might think of it [the ability to compare groups] as the place where instruction in the early years is headed and as the foundation from which further statistics will arise. Making such comparisons is the heart of statistics” (Konold & Higgins, 2003, p. 206).

Comparing Distributions: General Background

When discussing *comparing distributions*, this study adopts a statistical perspective. That is, we are neither comparing the size of two sets nor comparing attributes of two individuals, but rather comparing attributes from two groups with variability of elements in the groups (Konold et al., 1997).

Investigating students' understanding of single distributions has provided powerful insights into ways students come to view the relationship between isolated data points and their aggregate distribution (Bakker & Gravemeijer, 2004; P. Cobb, 1999; Konold & Pollatsek, 2002). Single distributions provide contexts for students to

investigate variability within a group only. According to the GAISE framework (Franklin et al., 2007), investigating variability within a group is appropriate for students at the beginning of the learning progression (Level A). In order to support students' reasoning about variability between two groups, as recommended for study by students at the next developmental level (Level B), *comparing distributions* becomes important. As Makar and Confrey (2004) discuss, *comparing distributions* provides motivation to learn statistics. In a context-rich environment, *comparing distributions* becomes a vehicle to support the investigation of centers and distribution; it provides the basis for hypothesis testing; and it allows variability to be considered in multiple ways. Other researchers, particularly those working with elementary and middle school students, have echoed the belief that *comparing distributions* provides a potentially fruitful reason for studying averages and thinking of variability (Bakker & Frederickson, 2005; P. Cobb, 1999; P. Cobb, McClain, et al., 2003; Konold & Higgins, 2003; Konold & Pollatsek, 2002; Petrosino, Lehrer, & Schauble, 2003; Watson & Moritz, 1999).

Prior research suggests that students have a tendency not to use averages to compare groups, even after significant experience with statistical ideas (Gal, Rothschild, & Wagner, 1989; Konold & Pollatsek, 2002). Konold and Pollatsek (2002) attribute the failings to students' missing the connection between the average and the entire distribution of values. They suggest instruction should encourage students to think of averages as measures of center or signals in noisy processes. In particular, they describe the use of repeated measures as potentially helpful because they can be seen as part signal and part noise more readily than other interpretations of average, such as fair share and

typical value. Further, they suggest it may be “fruitful to have students explore the relative stability of various indicators in different samples” (p. 192).

In order to thoroughly understand *comparing distributions*, an appreciation for variability is required (Bakker & Frederickson, 2005; Ben-Zvi, 2004; Makar & Confrey, 2002, 2004). Additionally, understanding other big statistical ideas such as randomness, bias, sampling, and measures of center are essential for making statistical comparisons (Franklin et al., 2007).

Comparing distributions in a dynamic software environment provides the opportunity to support the conceptual development of various measures of center and dispersion, to investigate characteristics of sampling distributions that are “grown” (Bakker & Frederickson, 2005) while looking for evidence with which to make decisions about relationships between groups. In addition to investigating within and between group variation, *comparing distributions* provides the opportunity to investigate ideas related to sampling distributions (Franklin et al., 2007; Makar & Confrey, 2004). Furthermore, Konold and Pollatsek (2002) suggest that designing and running controlled experiments may be useful for students learning to compare groups and to reason about average as a measure of center. As they said, “we expect that in a comparison situation, students can more easily view averages of the individual groups as summary measures of processes and can readily perceive the difference between those measures as some signal rising through the din of variability” (p. 285).

The following subsections within this section of the chapter explore student and teacher understanding as related to *comparing distributions*, distributions, variability, sampling distributions, and informal inference. Because the statistical content focus of

this study is on understanding *comparing distributions*, the first subsection is further broken into two detailed sections. The first section explores research on students' understanding of *comparing distributions* and is followed by a section about teachers' understanding of *comparing distributions*. The remaining sections combine research on student and teacher understanding of the statistical concepts of distributions, variability, sampling distributions, and informal inference. Research into teachers' understanding of statistical ideas is limited, so this study is informed by both research about student and teacher understanding.

Comparing Distributions: Student and Teacher Understanding

Comparing data sets or groups or distributions have emerged as viable statistical activities for students and teachers as well as a rich venue for research across age levels (Ben-Zvi, 2004; Lehrer & Schauble, 2004; Makar & Confrey, 2002; Pfannkuch et al., 2004; Watson & Moritz, 1999). Current research in this area covers the continuum from primary school to undergraduate studies to professional development work with pre-service and in-service teachers. Researchers have suggested that

most of the important issues and questions argued with data amount to comparing two groups, for example, treatment and control groups in medicine, before-and-after groups in various interventions and educational studies, and females versus males in gender equity studies. (Konold & Higgins, 2003, p. 207)

Research related to students' understanding of *comparing distributions* is presented next, followed by research related to teachers' understanding of *comparing distributions*.

Students' Understanding of Comparing Distributions

Calculating measures of center may interfere with reasoning from measures of center. One of the earliest examples of research involving student reasoning of *comparing distributions* was done by Gal et al. (1989). The study, involving 31 third graders and 31 sixth graders, found that students used a variety of strategies to compare two groups, but few students reasoned statistically. Even sixth graders who had studied the concept of average failed to use it when comparing two groups. In another study, Watson and Moritz (1999) investigated 88 grades 3-9 students' understanding of comparing two data sets using an individual interview protocol. They suggested that comparing two distributions can be motivating for students and should be part of the curriculum beginning at about grade 3 using equal-size data sets. They also reported the low proportion of students able to reason from the mean and suggested that, like Mokros and Russell (1995), prior experiences with the procedure to calculate the mean of a data set may interfere with students' ability to apply the concept appropriately. A number of other researchers have similarly reported learners', across ages, failure to use the mean or median to compare groups when it may have been appropriate (Bakker, 2004; Konold & Pollatsek, 2002; Rubin & Hammerman, 2006; Rubin, Hammerman, & Konold, 2006).

During a teaching experiment with middle school students, a research group discovered that though they had anticipated students would use the median when comparing two groups, students rarely did (P. Cobb, 1999). Instead, after exploring data representations using two specially-designed computer minitools and engaging in classroom argumentation regarding the interpretation of data, students began to use what

Cobb called, “modal clumps” or “hills” to compare groups, suggesting that students were beginning to reason with distribution rather than individual data points.

Reasoning with variability supports comparing distributions. Ben-Zvi (2004) detailed the work of two seventh-grade students in Israel as they began to develop views of variability in comparing two groups using tools to generate multiple statistical representations. In this case, students were comparing the length of surnames of 35 students in Israel with 35 students in the United States. He described seven stages through which the students progressed as they wrestled with variability within each group and then between groups, findings of which will be presented in the section on student and teacher understanding of variability. Experimenting with different tools and methods, previous experience with data sets, the context of the problem, the use of technology, and interactions with the teacher designed to help with the “negotiation of meanings” all apparently contributed to the students’ developing ideas of variability in *comparing distributions*. To support student reasoning as seen in this study, Ben-Zvi advocated for learning environments in which appropriate teacher guidance, peer work and interactions, and ongoing investigations with realistic problems are planned and managed. He acknowledged that the task of comparing data sets of equal size, though beneficial in this case, simplified the general case of comparing data sets of different sizes (p. 59).

Proportional reasoning is important when comparing two groups of unequal size. Watson and Moritz (1999) suggested a developmental progression through which students may move, which identified proportional reasoning strategies as essential for more sophisticated reasoning when comparing data sets of differing sizes. It is unclear

from this research, however, whether and in what ways a curricular influence may be confounding the developmental progression. As the sophistication of student responses increased generally with grade level, it is difficult to disassociate development with student opportunity to learn or general life experiences.

In an earlier teaching experiment reported by Cobb (1999), also with seventh graders, students were successful navigating comparisons of different-sized groups through their use of “hills” and eventually considered the proportion of data within various ranges of values, suggesting progress toward developing a statistical perspective.

Growing samples may support comparing distributions. Based upon classroom work with sixth-grade students, Bakker and Frederickson (2005) provided evidence to suggest *comparing distributions* and growing samples by hand and then with the assistance of technology such as *Tinkerplots* may enhance students’ understanding of measures of center, variation in and among distributions, variation in sample size and frequency, and the shape of data. Bakker and Frederickson recommend delaying the formalization of mean, median, and mode until students have had opportunities to consider variation and distribution in the context of data-driven investigations.

Graphical representations support and challenge students’ reasoning. Boxplots are useful representations when *comparing distributions* and are suggested as appropriate for students in grades 6-8 in the United States (NCTM, 2000). Bakker, Biehler, and Konold (2005) question whether grades 6-8 is an appropriate time to introduce this representation and argued, based upon teaching experiments with seventh-grade, eighth-grade, and university students, that boxplot representations are more challenging for

students to interpret than might be expected. Since boxplots display aggregate information rather than individual cases, and students often view data as individual cases, Bakker et al. (2005) recommend displaying dotplots and quartiles. Also since boxplots display relative density rather than frequency and the relative density is negatively related to the length of the component of the boxplot, they suggest use of a representation including both boxplots and dotplots simultaneously. Further they claim that medians pose difficulties for students to understand and recognize as a measure of center in the boxplot representation. Finally, because interquartile range (IQR) can be tricky due to ties in data values and differing software definitions for calculating IQR as a measure of spread, it may be difficult for students to see. Students tended to compute the five-number summary for boxplots and when comparing with another distribution, if all corresponding summary numbers were higher than the other set, they concluded one group had “larger values” and seemed to use a “shift model.” When these differences were not consistent, as is often the case, students had difficulty drawing conclusions.

In the same article (Bakker et al., 2005), a study by Biehler and Kombrink (1999) is referenced which suggests 15-year-old students who learned to sketch histograms from boxplots and boxplots from histograms, failed to learn to interpret the parts of the boxplot as measures of spread. Instead,

they regarded the median primarily as a cut point and not as a way to summarize where the data were centered. Furthermore, the students interpreted the interquartile range as “the spread of the middle half of the data,” rather than as a measure of spread that is a property of the whole data set. (p. 170)

The researchers referred to this interpretation as the “shape-summary” interpretation and compared it to the “center + spread” interpretation. When using some of these lessons in

a teaching experiment with university students, Biehler (1999) noted that students' success still required fairly direct instruction about statistical group comparison, with examples, counterexamples, and interpretations.

Teachers' Understanding of Comparing Distributions

Though statistics education is a blossoming research field, the research literature on teachers' understanding of statistical ideas is scarce (Ben-Zvi & Garfield, 2004a; Canada & Makar, 2006; Shaughnessy, 1992, 2007). Several studies on teachers' understanding of *comparing distributions* have been conducted by Makar and Confrey (Makar, 2004; Makar & Confrey, 2002, 2004, 2005). All of their studies involved pre-service or in-service teachers in statistical inquiry professional development or instruction. Other studies involving teachers and *comparing distributions* have been conducted during professional development surrounding teachers' use of *Tinkerplots* and *Fathom2* software to explore and analyze data (Hammerman & Rubin, 2006; Rubin & Hammerman, 2006; Rubin, Hammerman, & Konold, 2006).

Secondary mathematics teachers' professional development. One study was situated at the end of a six-month professional development sequence designed to support secondary teachers' interpretation of state-mandated student test data (Makar & Confrey, 2004). This study involved four secondary mathematics teachers (two middle school teachers, one pre-service teacher, and one high school teacher) and the ways in which they decided whether two groups differed. The professional development encouraged teachers to engage in statistical activity as investigators and used simulations with *Fathom2* software to explore sampling distributions and other statistical content. A pre-

post content test indicated significant growth in the areas of sampling distributions, inference and hypothesis testing, and overall. To better understand teachers' statistical reasoning when *comparing distributions*, qualitative analysis resulted in four final categories: measurable conjectures, tolerance for variability, understanding of context, and a view towards inference. These categories were used as the basis for the development of the framework (Table 1) for examining statistical reasoning when *comparing distributions* (Makar & Confrey, 2002).

The teachers in Makar and Confrey's (2004) study reasoned about variability when *comparing distributions*, in three different ways: (1) as variation within a group—the variability of data; (2) as variability between groups—the variability of measures; and (3) distinguishing between these two types of variability. Teachers in this study recognized variation within a group, but struggled to quantify variation between distributions. As a consequence, the authors recommended that sources of variation in data and measures be discussed frequently when working with data and again as measures are compared between distributions to develop a tolerance for variation both within and between distributions. Additionally, Makar and Confrey recommend the use of simulations including the randomization test, to support the conceptual development of concepts related to an inference-like view of a difference between two groups, but caution that without care this approach may promote misconceptions of sampling distributions. Because of the small number of teachers willing to participate in this study, one of the issues raised by Makar and Confrey was the difficulty of engaging secondary teachers in research designed both to influence and study teacher learning and practice.

Table 1

Taxonomy for Classifying Levels of Reasoning When Comparing Two Groups

Level 1	Level 2	Level 3	Level 4	Level 5
Pre-descriptive	Descriptive	Emerging Distributional	Transitional View	Emerging Statistical
No recognition of relationships between datasets is made, except based on individual data points or anecdotal evidence. If conjectures are made at this level, they are unmeasurable.	Focus on summary statistics and make absolute comparison between datasets with no regard for variability. Conjectures assume data is infinitely available to answer any question.	First holistic view of the data; informal qualitative descriptors of the data, along with basic summary statistics, are used to describe two datasets. Teachers begin to understand the difficulty in creating measurable conjectures, but are unable to successfully resolve the conflict and show frustration in attempting to write an appropriate conjecture. Variability, while acknowledged, is not understood beyond a descriptive level.	Begin to understand the influence of variability in comparing two groups. More flexibility is shown (e.g., multiple graphical representations, alternative measures of center or spread) in comparing datasets at this level. Conjectures, while questionably measurable, have progressed to show elementary understanding of the difficulty in creating a conjecture that doesn't overly compromise the question at hand, but allows for possible collection of data.	Gain confidence in using standard descriptive statistics to compare data sets, taking into consideration the differences between measures of center in light of the variability in the data and the sample sizes of the datasets. Conjectures demonstrate some ability to frame questions that balance data constraints with the problem at hand. Context and quantified description are well integrated into conclusions and inferences may attempt to draw on statistical models, if relevant

In another publication referring to the same study, Makar and Confrey (2002) introduced the taxonomy in Table 1 for classifying reasoning when comparing two groups. While using this taxonomy to categorize teachers' reasoning, pre-post test results showed improvement; however, when the taxonomy was applied to results from teachers' individual statistical investigations, two of the four teachers regressed considerably on the taxonomy.

An important consideration in using or adapting Makar and Confrey's (2002) framework is that it was generated during a professional development intervention with a small group of teachers. Initially 11 teachers participated in the professional development; however, by the end of the session, only 4 remained. Small sample sizes have been a hallmark of statistics professional development research and reinforce the need for larger studies such as the present one (Shaughnessy, 2007).

Pre-service undergraduate students. Makar's (2004) dissertation examined prospective secondary mathematics and science teachers' understanding of equity, fairness, and statistical inquiry as they learned to analyze accountability data. *Comparing distributions* was not a focal point of her study, but one of her research questions that related to the present study was "What level and types of understanding of the concepts of distribution and variation were learned?" In order to answer that question, she conducted a pre-post content knowledge test of 18 pre-service teachers. The items and related results from her study were influential in the design of the pre-post instrument for the present study. Specific items, results, and comparisons will be provided in Chapter V. Two other findings from Makar's study relevant to the present study are (1) all participating teachers had developed a distributional view of data through their explorations, and (2) teachers' informal language use when referring to variability (e.g., spread, cluster, clump) is an important consideration in supporting understanding (Makar, 2004; Makar & Confrey, 2005).

Based upon a synthesis of the literature presented thus far, it is likely that understanding *comparing distributions*, at a level appropriate for high school teachers,

may involve understanding of distributions, variability, sampling distributions, and informal inference. Literature related to each of these sub-ideas will be presented next.

Students' and Teachers' Understanding of Distributions

The idea of “distribution” as a unifying concept in statistics has emerged through the work of various researchers (Bakker, 2004; P. Cobb, 1999; McClain, Cobb, & Gravemeijer, 2000). Using design experiments and computer microworlds, researchers have studied the ways in which students come to view distributions as aggregates rather than collections of individual cases. An important finding from this work is that students often use informal language as they begin to reason about distributions, for example, words like “clump,” “bump,” “majority,” “outlier,” “reliability,” and “spread out” were used as students attempted to explain distribution characteristics.

Bakker and Gravemeijer (2004) suggest that informal reasoning about distribution may be based upon reasoning with shapes. They relate data (individual values) to distribution (conceptual entity) through concepts of center, spread, density, and skewness, each with a close connection to shape. Using specially-designed Minitools with seventh graders to conduct statistical explorations and design research methodology, students progressed through three stages of tool use: Stage 1, data are represented by bars; Stage 2, dots replace bars; and Stage 3, symbolizing data as a “bump.” Bakker and Gravemeijer claim that students should be allowed to create personally meaningful representations for data and to use language which may lack statistical precision as they endeavor to make sense of and communicate their statistical understandings. Furthermore, they recommend students have experiences with both interpreting graphs and constructing graphs with special statistical properties in order to fully coordinate ideas of distribution.

Bakker and Gravemeijer (2004) recommend three heuristics to support student learning to reason about distributions: (1) letting students invent their own data sets could stimulate them to think of a data set as a whole instead of individual data points; (2) growing samples is a promising instructional activity to let students reason with stable features of variable processes; and (3) predictions about the shape and location of distributions in hypothetical situations are useful to foster a more global view and to let students see the signal in the noise. Of particular importance to the present study is the suggestion that big ideas of sampling and distribution may be developed coherently, presumably related to heuristic 2. “Without variation, there is no distribution, and without sampling there are mostly no data. We therefore chose to deal informally and coherently with all these big ideas at the same time with distribution in a central position” (p. 149).

Konold and Pollatsek (2002) argue that the signal within the noise metaphor is useful for students to understand distribution. In this way, measures of center represent the signal within the noise of individual data points and the concept of distribution is developed as “distribution around” the signal (p. 171). Mokros and Russell (1995; Russell & Mokros, 1990) found that fourth-, sixth-, and eighth-grade students struggled with conceptions of the mean of a distribution. They found students reasoned about the average as mode, algorithm, reasonable, midpoint, and balance point, with those reasoning from the perspectives of the mode or the algorithm experiencing more difficulty articulating a well-constructed and articulated idea of the average, that is, not viewed as representative of the distribution.

In a study with 17 secondary pre-service teachers, Makar and Confrey (2005) documented teachers’ use of informal language when describing variation and

distributions and suggested the need for teachers to recognize and value “variation talk” in order to encourage statistical sense-making. Like other researchers (e.g., Bakker, 2004; P. Cobb, 1999; Konold & Higgins, 2003), Makar and Confrey’s research suggests the need to allow students and teachers to develop and communicate intuitive ideas about distributions and variability prior to the use of more formal definitions. Informal language becomes important in negotiating shared meanings in a classroom setting and also when considering assessment issues. Because understanding may be masked by lack of formal language acquisition, students and teachers may have developing conceptions of statistical ideas that will fail to be communicated or recognized in reference to standard statistical language.

Meletiou-Mavrotheris and Lee (2003) found that even after instruction, undergraduate students from an introductory class in which understanding of histograms and other graphical tools were stressed, had difficulty constructing and interpreting histograms and confused graphical representations. A number of other researchers have reported students’ and adults’ difficulty with constructing and interpreting boxplots; the inability to discriminate between bar graphs and histograms has been seen repeatedly across studies (e.g., Baker, Corbett, & Koedinger, 2002).

Students’ and Teachers’ Understanding of Variability

Research in understanding the development of the concept of variability in students has been said to be embryonic (Watson et al., 2003). It has been within the last 10-15 years that research efforts have turned to issues of variability and much of this recent work has focused on either middle school students’ or college students’

conceptions of variability (e.g., Bakker, 2004; Meletiou-Mavrotheris & Lee, 2002, 2003; Mickelson & Heaton, 2003; Watson et al., 2003). A particular focus has been on assessing students' conceptions of variability and attempting to create frameworks designed to capture and quantify that understanding (Canada, 2006; Makar & Confrey, 2002; Mooney, 2002; Reading, 2004; Watson et al., 2003). It has been suggested that research typically mirrors the emphasis placed in curricular materials, which have historically lacked a focus on variation (Reading & Shaughnessy, 2004; Torok & Watson, 2000). The lack of research into student and teacher understanding of variability until recently might be explained because of the lack of curriculum materials attending to the concept of variability.

According to Wild and Pfannkuch (1999), consideration of variability is a hallmark of statistical thinking and includes the following components: noticing and acknowledging; measuring and modeling for the purposes of prediction, explanation, or control; explaining and dealing with; and investigative strategies. Consideration of variability is one piece of their four-dimensional framework representing statistical thinking in empirical inquiry. The development of the framework was based upon students' work on statistical tasks; interviews with team leaders of statistical projects; and interviews with six practicing statisticians. If one assumes that the model accurately portrays the dimensions and components of statistical thinking, it may be useful for planning learning opportunities and environments for students and teachers.

Students' reasoning with variability. With appropriate tools and teacher support, fourth-grade students have demonstrated unusually sophisticated reasoning based on the concept of distribution and its relationship to the distribution of error terms resulting from

experimental measurements (Petrosino et al., 2003). In this study with 22 students, Petrosino et al. (2003) found that students could jointly consider center and spread when *comparing distributions* in measurement contexts. The measurement process was seen as important in order for the students to consider variation between measurements and within a distribution. During the course of the study, students measured rocket height, pencil length, and the height of a flagpole, using tools designed to increase or decrease variability in measurements in order to highlight differences and to support comparisons between and among measurements. Students were able to argue convincingly that one rocket design outperformed another, even when the experimental results contradicted their original hypothesis. Analysis and results of students' additional written and interview tasks indicated that student performance was impressive for students at fourth-grade level, but even more impressive when compared to that of older students. Given the success of these fourth graders and their ability to reason in statistically powerful ways when supported with appropriate tools and teacher guidance, lessons from this study informed the design of investigations used in the present study.

As described in the section devoted to student understanding of *comparing distributions*, two seventh-grade students in Israel learned to reason about variability through an investigation involving comparing the lengths of students' surnames from Israeli and American classrooms (Ben-Zvi, 2004). These students were provided access to Excel software and the two groups were of equal size. Ben-Zvi documented seven stages through which these able and verbal students progressed as they participated in the surname investigation: (1) On what to focus: Beginning from irrelevant and local information; (2) How to describe variability in raw data; (3) How to formulate a

statistical hypothesis that accounts for variability; (4) How to account for variability when comparing groups using frequency tables; (5) How to use center and spread measures to compare groups; (6) How to model variability informally through handling outlying values; (7) How to notice and distinguish the variability within and between the distributions in a graph (p. 48). A carefully designed learning environment supported the progress made by these students. In particular, the curriculum embodied EDA through its use of semi-structured, open-ended, extended meaningful problem contexts, teacher as a non-directive guide, computerized tools for handling the complexity of calculations and representations, and conceptual discussions.

Student conceptions of variability include variability (1) in particular values, including extremes or outliers; (2) as change over time; (3) as whole range; (4) as the likely range of a sample; (5) as distance or difference from some fixed point; (6) as the sum of residuals; (7) as covariation or association; (8) as distribution (Shaughnessy, 2007). Because of the numerous ways in which variability can be manifested, for example in data, in samples, and in distributions, many opportunities for students to learn about or manifest difficulty with variability exist. Teachers should be acquainted with each of these conceptions of variability in order that their statistical knowledge for teaching is adequate.

In a large study with school age students, an assessment instrument to measure students' understanding of variability was designed and validated using a Rasch analysis (Watson et al., 2003). The hierarchy of levels of understanding variability in Table 2 was a result of that research and incorporated variability reasoning using basic chance items,

basic tables and graphs, variation in chance items, variation in data and graphs, and variation in sampling items.

Table 2

Framework for Levels of Understanding of Variability

Level 1	Level 2	Level 3	Level 4
Prerequisites for variation	Partial recognition of variation	Applications of variation	Critical aspects of variation
Working out the environment, table/simple graph reading, intuitive reasoning for chance	Putting ideas in context, tendency to focus on single aspects and neglect others	Consolidating and using ideas in context, inconsistent in picking salient features	Employing complex justification or critical reasoning

This framework was used in the present study to help clarify teachers' level of understanding of variability. Though not designed for characterizing teachers' understanding, the levels appear appropriate for the purposes of the present study. The coordination of this framework with other frameworks for assessing teachers' understanding as it relates to *comparing distributions* will be discussed in more detail in Chapter III.

Pre-service teachers' reasoning with variability. In a study with pre-service teachers, Canada (2006) investigated the effects of hands-on activities, computer simulations, and discussions designed to support teachers' attention to variability. The results of the study suggested that these components contributed to teachers' enhanced attention to variability in a probability environment. Canada elaborated a framework to represent teacher understanding which included three main aspects: (1) expecting variation, (2) displaying variation, and (3) interpreting variation, each with two dimensions. What is highlighted in the study is the need for points of reference for comparing qualitative responses from an environment where such responses are

encouraged. Canada's work, though in the context of elementary pre-service education, supports the need for careful study of the ways in which teachers and students think about statistical topics, thus providing direction for potential ways to influence that thinking through classroom activity.

Framework for instruction and assessment of variability. Garfield and Ben-Zvi (2005) introduced seven increasingly sophisticated areas of knowledge of variability that may be useful for instruction and assessment: (1) developing intuitive ideas of variability, (2) describing and representing variability, (3) using variability to make comparisons, (4) recognizing variability in special types of distributions, (5) identifying patterns of variability in fitting models, (6) using variability to predict random samples or outcomes, and (7) considering variability as part of statistical thinking. They state, "ideas related to variability must be constantly revisited along the statistics curriculum from different points of view, context and levels of abstraction, to create a complex web of interconnections among them" (p. 95).

Collectively, the literature in this section suggests that variability is an essential component in statistical thinking and reasoning, is accessible to even elementary grade students under appropriate conditions, and its conceptual development is complex and requires repeated experiences under varying conditions.

Students' and Teachers' Understanding of Sampling Distributions

A number of researchers have studied issues surrounding the difficulty students have in learning about sampling distributions (Chance et al., 2004; G. W. Cobb & Moore, 1997; delMas et al., 1999; Saldanha, 2004; Saldanha & Thompson, 2002). Related to the

understanding of sampling distributions is that of sampling. Recently a number of researchers have studied school-age students' reasoning in sampling situations across age-levels and have found sampling reasoning to be challenging for students (Bakker & Frederickson, 2005; Rubin, Bruce, & Tenney, 1990; Saldanha & Thompson, 2002; Shaughnessy, Ciancetta, & Canada, 2004; Watson, 2004; Watson & Moritz, 2000). Many students tend to predict samples with inappropriate distributions or reason on the basis of small or non-random samples, given the context of problems. This research augments the seminal psychological research of Tversky and Kahneman (Kahneman & Tversky, 1972; Tversky & Kahneman, 1974), in which they identified a number of persistent misconceptions, heuristics, and biases people used when reasoning in probabilistic situations. One of the heuristics they identified was that of representativeness.

Sample representativeness is the idea that a sample taken from a population will often have characteristics similar to those of its parent population. . . . Sample variability is the contrasting idea that samples from a single population are not all the same and thus do not all match the population. (Rubin et al., 1990, p. 314)

When students rely too heavily on sample representativeness, they tend to believe the sample tells us everything about a population; whereas when students rely too heavily on sample variability, they tend to believe the sample tells us nothing about a population (Rubin et al., 1990). Reasoning somewhere between the two extremes may be a reasonable target and perhaps experiences including physical and simulated sampling may be beneficial to learners of statistics in developing an appreciation of the power of sampling and its usefulness to the formulation of statistical arguments.

In a study of undergraduates in an introductory statistics class designed to engage students in statistical activity and to present statistical thinking as a balance between

deterministic and stochastic reasoning, students did show improved understanding of the relationship between chance and regularity and improvements in skills and dispositions (Meletiou-Mavrotheris & Lee, 2003). However, “Most students in the class failed to grasp the relationship between population, sample, and sampling distribution and, consequently, between population standard deviation, sample standard deviation, and standard error” (p. 46).

Saldanha and Thompson (2002) discuss the distinction between the way that statisticians use results from a sample to make inferences and the need for students to develop an appreciation of why statisticians are confident doing so. They distinguish between two conceptions of “sample” that occurred during a teaching experiment with 27 11th- and 12th-grade students and suggest that “sample as quasi-proportional small scale version of the population” (p. 2) is a better target for instruction than the alternative “sampling as a subset of the population” view. In this quasi-proportional small scale scheme, the multiplicative conceptual structure (MCS) of distribution is supported and issues of variability and resampling become relevant in the construction of sampling distributions.

The use of simulations to support student and teacher learning of sampling distributions has been seen throughout the recent literature (e.g., Chance, Garfield, & delMas, 2000; delMas et al., 1999; Konold, 1994; Lunsford, Rowell, & Goodson-Epsy, 2006; Wood, 2005). Some research relating use of simulations to learning about sampling distributions is presented here; a more thorough review of the use of simulations and resampling methods is presented in a later section of the chapter.

Konold (1994) designed simulation-based activities for students to explore sampling distributions through use of the randomization test. Although Konold acknowledges that students appeared to appreciate the opportunity to learn probability and statistical inference using resampling approaches, on post-intervention assessments, students demonstrated surprisingly weak understanding of the probabilistic ideas underpinning their work with sampling distributions. These poor results suggest the need for better understanding of sampling distributions and the mechanisms which might support learners' conceptual understanding. Disappointing results from Konold were later followed by those of delMas et al. (1999), confirming the difficulty in developing understanding of sampling distributions and suggesting that use of computer simulations may encourage the development of misconceptions, rather than advance understanding of sampling distributions.

delMas and colleagues have continued to explore college students' understanding of sampling distributions in a simulation environment, perhaps because like others, they believe that simulations should support student understanding even though empirical results have been lackluster (Chance et al., 2004; Saldanha, 2004; Saldanha & Thompson, 2002). After extensive research on students' understanding of sampling distributions, Chance et al. (2004) recommend: (1) Use the technology to first explore samples and compare how sample behavior mimics population behavior; (2) Provide students with the experience of physically drawing samples; (3) Allow time for both structured and unstructured explorations with the technology; (4) Discuss the students' observations after completing the activity; (5) Repeatedly assess student understanding of sampling distributions; and (6) Build on students' understanding of sampling distributions later in

the course. These recommendations are made in the context of a study of undergraduates in an introductory statistics course with access to simulation software. These recommendations are hypothesized to be applicable to the study of sampling distributions for other adult learners.

On the basis of an extensive teaching experiment (Saldanha, 2004) with eight high school students, developing the concept of sampling distribution, even in a specially-designed, simulation-inspired environment, appears to be non-trivial. In particular, Saldanha suggests that it may be the hierarchically structured objects and processes involved in the construction of sampling distributions which prove problematic for students. “This points to a potentially relevant area for further research: investigating the conceptualization of hierarchically structured objects and processes” (Saldanha, 2004, p. 268).

The three-tier Simulation Process Model (SPM) is a graphic organizer representing the explicit connections between the hierarchically structured objects involved in the simulation of sampling distributions that may be useful in supporting students’ understanding of sampling distributions (Lane-Getaz, 2006). In the SPM, the first tier refers to the population, the second tier refers to the samples or statistics from the sample, and the third tier represents the distribution of sample statistics. This model generally captures the logic of inference from posing the “What if . . .?” question, to generating random samples of size n , to selecting a summary statistics, to compiling the summary statistics, and finally to assessing the rareness of the observed sample statistic relative to the distribution of sample statistics from the simulation. Lane-Getaz (2006)

speculates that the use of this pre-organizer will help to scaffold students' thinking from informal inference to more formal tests of significance.

The SPM also models the connections between the hierarchically-structured objects in the Randomization Distribution in *CPMP-Tools* as well as the structure one may construct in *Fathom2* to simulate the randomization distribution or other sampling distribution. It is this transparency of structure that was influential in the design of the professional development intervention for the present study.

Through continued work with college students, a framework for describing students' statistical reasoning about sampling distributions (Table 3) was developed by Chance, delMas, & Garfield (2004, pp. 302-303).

Table 3

Stages of Development in Students' Statistical Reasoning about Sampling Distributions

Level 1	Level 2	Level 3	Level 4	Level 5
Idiosyncratic Reasoning	Verbal Reasoning	Transitional Reasoning	Procedural Reasoning	Integrated Process Reasoning
Student knows words and symbols related to sampling distributions, uses them without fully understanding them, often incorrectly, and may use them simultaneously with unrelated information.	Student has a verbal understanding of sampling distributions and implications of the Central Limit Theorem, but cannot apply this to the actual behavior of sample means in repeated samples.	Student is able to correctly identify one or two characteristics of the sampling process* without fully integrating these characteristics.	Student is able to correctly identify the three characteristics of the sampling process* but does not fully integrate them or understand predictable long-term process.	Student has a complete understanding of the process of sampling* and sampling distributions, in which rules and stochastic behavior are coordinated.

Note: *A total of three characteristics are identified: mean of sampling distribution is equal to population mean; shape of sampling distribution becomes more normal as sample size increases; variability of sampling distribution decreases as sample size increases.

When trying to validate the levels of reasoning in Table 3 during a follow-up study that included extensive video-tape analysis, Chance et al. (2004) experienced

difficulty clearly placing students into categories. They suggested additional dimensions of statistical reasoning that may be important to consider because of the complexity of students' reasoning. These dimensions included students' fluency with terminology, concepts, and procedures; the degree to which students identified and used formal rules for predictions and explanations; students' consistency of reasoning; students' acknowledgement of inconsistencies in reasoning; and student confidence in reasoning. The present study recognized difficulties in categorizing students' reasoning using this framework and was mindful of the additional dimensions, but also coordinated the framework with two other frameworks with the intent of facilitating an accurate categorization of teachers' reasoning when *comparing distributions*. This coordination is further discussed in Chapter III.

Given that students as early as fourth or fifth grade have been able to reason in sophisticated ways regarding sampling distributions within a carefully constructed learning trajectory (Lehrer & Schauble, 2004), it is reasonable to anticipate that under appropriate conditions, adult learners should be able to come to understand sampling distributions. Based on the literature, sampling distributions are particularly difficult for students to comprehend; thus, more studies are needed to shed light on how this understanding may be more carefully supported. Understanding sampling distributions may then provide one more rung for students on the ladder to understanding informal inference.

Informal Inference: What Is It? and What Do Students Know?

Garfield and Ben-Zvi (2004) suggest inference includes the way of estimating and drawing conclusions about larger groups based on samples. In inferential statistics,

the question is: Given the data, what can we say about specific aspects of the stochastic mechanisms that governed the occurrence of those data? The actual data are regarded as a result of a random process, in the sense that if the data collection were to be repeated, the outcome would typically be different. Consequently, whatever inference we make from the actual data, it is subject to error. This error—the central concept of statistics—is not meant to be a ‘mistake’ of any kind (i.e., something that could be avoided). It refers to the unavoidable randomness: If the data were collected again, we might have reached a different conclusion. (Bartoszynski & Niewiadomska-Bugaj, 1996, p. 442)

Statistical inference has been seen as difficult for students and requires a good deal of mathematical maturity to handle the mathematical demand of the development of theory to support inference (Ben-Zvi & Garfield, 2004a; Meletiou-Mavrotheris & Lee, 2003; Moore, 1997). As an example, a correct interpretation of p -values is essential for communicating the results of statistical tests used for inference (e.g., t -tests, chi-square tests, ANOVA). In an experimental study on the understanding of significance, Krauss and Wassner (2002) found that when they presented college-level psychology students ($n = 44$), scientists not teaching methods courses ($n = 39$), and statistics methods teachers ($n = 30$) with six statements representing common misconceptions of the interpretation of a significant test result, more than 80% of each group erroneously classified at least one of the statements as correct. What was especially surprising in this study is that 80% of the people teaching statistical methods exhibited the commonly held misconceptions. There are a number of possible explanations for this phenomenon; however, one of the possible courses of action to help improve inferential thinking and reasoning may be to

support the development of informal inference prior to the study of more formal techniques and theory of inference.

While working on the VISOR (Visualizing Statistical Relationships) project with middle and high school mathematics teachers, Rubin et al. (2006) tried to support teachers' understanding of informal inference. They defined informal inference as reasoning that involves the following related ideas of properties of aggregates, sample size, controlling for bias, and tendency. Within properties of aggregates, they consider signal and noise, and types of variability. Specific types of variability include variability due to errors of measurement, multiple causes, and sampling. During the VISOR project investigations, teachers used *Tinkerplots* (Key Curriculum Press, 2005b) software to explore relationships within data sets in ways that would not be feasible without the flexible technology (Rubin et al., 2006). Using technology and a richly constructed cover-story, three middle and six high school teachers successfully argued for a solution to a problem which required sophisticated statistical reasoning with heavy emphasis on signal and noise.

The ideas of informal inference explicated by Rubin et al. (2006) have direct connections to the present study. With similar goals, the present study supported teachers' informal inference through the exploration of tasks requiring statistical reasoning and argumentation and the use of technology for creating and managing multiple representations. The learning activities, which will be discussed in Chapter IV, were designed with potential to encourage the development of strong connections to each of the related ideas associated with informal inference as outlined by Rubin et al. (2006).

According to Pfannkuch et al. (2004), Biehler (2001) discussed a four-stage development process for learning formal inference. Biehler's stages when comparing two boxplots would include (1) fine tuning the comparison, (2) widening and exploiting the context by bringing in more variables, (3) generalization, and (4) can the group differences be due to chance? Research on the last stage of Biehler's model is limited (Pfannkuch et al., 2004). Because resolving stage four relies on students' reasoning with sampling distributions, this is an area in which research overlaps. As seen in the section on sampling distributions, it is yet unclear as to what remedies may help students effectively grapple with the complexities of sampling distributions. Simulations and resampling techniques have been utilized in several studies and the general consensus is that they have merit for helping students develop understanding of sampling distributions, but empirical evidence is yet not strong. The present study extends this line of research.

From a study involving two secondary teachers attempting to teach informal inference to students, Pfannkuch et al. (2004) suggested a pedagogical framework to inform teachers of the big ideas students need to understand inference: "(i) knowing why they should compare centres, (ii) describing and interpreting variability within and between sample distributions, (iii) developing their sampling reasoning, and (iv) how to draw an acceptable conclusion based on informal inference" (p. 5). Pfannkuch and colleagues purport that "Without attention to the complexity of informal inference and to the provision of a teaching pathway towards formal inference, statistical inferential reasoning will continue to elude most students" (p. 7).

From a separate study involving 15-year-old students and a teacher in New Zealand, Pfannkuch (2006) states, "Informal inferential reasoning is interconnected to

reasoning from distributions, reasoning with measures of centre, and sampling reasoning within an empirical enquiry cycle. All of these aspects are underpinned by a fundamental statistical thinking element, consideration of variation” (p. 1). She argues that prior to formal inference, much of the work of drawing conclusions from data relies on evaluating graphical representations. In her study, she was concerned with informal inference involving comparison of boxplots. She described 10 different views used by the teacher of these students to compare boxplots during a teaching experiment: (1) hypothesis, (2) summary, (3) shift, (4) spread, (5) signal, (6) sampling, (7) explanatory, (8) individual case, (9) referent, and (10) evaluative. It is important to note that Pfannkuch’s research subjects did not have access to technology to support their reasoning. Though one-third of the students reasoned from a signal view, it is still unclear how instruction can reinforce this productive disposition. Also only 4 of 29 students were found to reason at the highest level, Level 3, suggesting that student progress was somewhat limited.

Overall, results suggest that

improvement in inferential reasoning may depend upon more awareness of the multiple views taken when reasoning with boxplots, developing teacher and student talk, keeping data under the boxplots for as long as possible, and giving more opportunities to students to experience sampling behavior. (p. 6)

How student use of technology may have influenced this finding is an open question.

Technology to Support Statistics Learning

Fathom2 (Key Curriculum Press, 2005a) and *CPMP-Tools* (Keller, 2006) are statistical learning tools which may provide improved environments for students and teachers to conduct simulations over those used in previously reported research. *CPMP-*

Tools is Java-based software which allows users to investigate existing datasets as well as to create their own data sets of interest. *CPMP-Tools* is appealing because of its user-friendly interface and because it was designed with specific curricular goals, although it does not have all of the features of a full-blown statistical package. It would likely be characterized by Bakker (2002) as route-type software because of its fairly direct and relatively inflexible design. An affordance of this tool is that users can easily navigate between the limited features of the environment, and thus the technology supports student investigation rather than creating a steep learning curve toward successful operation. The interface has been designed to highlight the hierarchy of the structured objects and processes to support learning, similar to the SPM described earlier. The ease with which students can navigate between representations and the dynamic animation supports exploratory data analysis (EDA) and increases the likelihood that students will be able to engage with the technology. *CPMP-Tools* may be a useful resource for students to investigate characteristics of distributions and compare two distributions. Of particular interest in this study are the *CPMP-Tools*' Randomization Distribution feature and Balancing Histogram and Estimating Standard Deviation custom tools.

Fathom2 is a dynamic statistical software package designed in the spirit of EDA and with built-in features making the tool incredibly flexible and useful for investigations. Bakker (2002) would refer to this as a landscape-type tool because of the flexibility inherent in the design and use of the tool. The basic structure of a simulation in *Fathom2* parallels that which one would conduct by hand and the hierarchical nature of the process can become visible, parallel to that described in the SPM. The time required to navigate in this system is significantly greater than that with *CPMP-Tools*; however, it

is conjectured that the added learner flexibility may be worth the effort. Furthermore, the structure of objects in *Fathom2* may support understanding of *comparing distributions*.

Using any new tool or representation necessitates the change in the content and pedagogy of statistics instruction and in many cases teachers are unprepared for these changes. . . . Professional development for teachers will need to address issues of mathematical content, as well as issues of learning, representation, and pedagogy. By exploring and discussing data for themselves in new ways, teachers can develop a deeper understanding of the mathematics, and also of how classroom discourse and pedagogy might change through use of new software tools. However, *teachers'* experiences of learning with these new software tools have not yet been explored. (Hammerman & Rubin, 2004, pp. 18-19)

Though careful study of students' or teachers' use of *CPMP-Tools* has not been documented and research around the use of *Fathom2* is just beginning to emerge, it is likely that the use of these tools will affect the ways in which teachers come to understand statistical ideas and representations. For the purposes of the present study, these tools were utilized most heavily through the conduct of simulations during statistical investigations.

Use of Simulations

With the ever-increasing power and availability of computers, methods of utilizing the computing and processing capability of the machines have been devised to simulate complex mathematical and statistical situations. The use of computer technology for purposes of simulation can be found in research literature back more than 30 years. Though researchers believe simulations to be powerful pedagogical tools to support student understanding, little empirical evidence exists to support the claim (Mills, 2002). Still, researchers have generally recommended using computer simulation methods (CSM) to teach statistics concepts that are particularly difficult or abstract (Mills, 2002).

A review of the literature since 1983 regarding teaching with CSMs at the post-secondary level suggests that CSMs have been used for teaching concepts of the Central Limit Theorem, the t -distribution, confidence intervals, the binomial distribution, regression analysis, sampling distributions, hypothesis testing, and survey sampling. Though researchers report successes in teaching, there is very little empirical and theoretical research to substantiate the recommendations that use of simulations improve student learning (Mills, 2002). Two results from empirical studies mentioned in this review are noteworthy: (1) CSMs appeared to be effective for lower-ability students, and (2) learning appears to be enhanced when CSMs are used to provoke students to confront their faulty ideas or misconceptions (delMas et al., 1999; Mills, 2002).

A line of related research which appears absent from Mills' review of CSMs is research involving Monte Carlo simulations. Prior to 1983, the use of Monte Carlo simulations in teaching was seen as a promising approach to statistics instruction (Atkinson, 1975; Hecht, 1980; Shevokas, 1974; J. L. Simon, Atkinson, & Shevokas, 1976). Three teaching experiments at the University of Illinois utilizing Monte Carlo simulations to teach probability and statistics suggested that the Monte Carlo approach was viable and preferable for teaching probability concepts (J. L. Simon et al., 1976).

Simon and colleagues suggested that

The Monte Carlo method is not offered as a successor to analytic methods. Rather, it can be an underpinning for analytic teaching to help students understand analytic methods better. . . . The Monte Carlo method is not explained by the instructor. Rather it is discovered by the students.
(p. 734)

Furthermore, "an average university class can be brought to re-invent such devices as the Monte Carlo version of Fisher's randomization test" (p. 735). The systematic Monte

Carlo method at the University of Illinois was defined as: (1) Construct the universe whose behavior one is interested in; (2) draw a sample from that universe; (3) compute the statistic of interest; (4) repeat the sampling procedure a large number of times; (5) calculate the proportion of “successes” to experimental trials, which estimates the probability of the event in which one is interested.

Since that time, Monte Carlo approaches have been investigated further to suggest potential for the support of student learning in probability and statistics (J. L. Simon, 1997; Wagner-Krankel, 1990). For example, Wagner-Krankel (1990) found that below-average ninth-grade students in a suburban Chicago public high school who studied probability using Monte Carlo techniques learned more probability than did students from the comparison group without such opportunity. Because the classes were general mathematics and consumer mathematics environments, it is unclear whether students in the comparison group had the opportunity to study probability to the same extent as the experimental treatment group; however, given that more students scored 70% or above on the final exam from the treatment group than the control group, the authors make the case that experience with Monte Carlo techniques supported improved student learning (Wagner-Krankel, 1990).

A study with 40 undergraduate students in an introductory statistics course suggested that the use of simulation software increased student achievement and beliefs suggestive of the value of simulation (Sterling & Gray, 1991). In a discussion of the merits of the using simulations to support learning, Lane and Peres (2006) suggest that “Guided discovery learning in which students are asked questions before they interact with the simulation and then use the simulation to confirm or disconfirm their answers

appears to be an effective technique” (p. 5). Lane and Peres refer to this technique as “query first.” In a sense, students were making predictions and then testing them to see whether they were correct, thus invoking the possibility of cognitive conflict or disequilibrium. Use of simulations in this way avoids the problem of students simply being passive observers during a demonstration lesson by virtue of their minds being engaged. Studies in which simulations were used and students were either not actively engaged or were provided too much direction show little to no positive learning affect on students (e.g., delMas et al., 1999).

Use of Resampling Methods

Like Simon, this study adopts the position that,

“Simulation” here means any process of Monte Carlo experimental repetitions of a model of a probabilistic process. “Resampling” refers to the subclass of statistical problems done with simulations; it includes bootstrap and Fisherian permutation methods as well as simulation techniques for dealing with a variety of other problems such as binomial proportions. (J. L. Simon, 1994, p. 290)

Resampling refers to the use of observed data or of a data generating mechanism (such as a die or computer simulation) to produce new hypothetical samples, the results of which can then be analyzed (J. L. Simon, 1997).

Resampling methods were introduced in the 1930s, but were quickly displaced by less powerful, less accurate techniques and approximations that required the use of tables based upon theoretical calculations (Good, 1999). The lack of computing capability to handle extensive data sampling and calculations was not available at the time. With the cheap and available power of desktop computers today, resampling methods have become viable and recognized as powerful means of supporting statistical thinking and

reasoning of students (Edgington, 1995; Good, 1999). Resampling methods provide the added benefit that mathematics beyond the level of high school algebra is not required to understand the intuition of the methods. Resampling methods include specific techniques of randomization testing, bootstrapping, and the jack-knife.

Permutation tests, also called randomization tests, fall under the category of resampling methods and produce results whose distribution approximates a sampling distribution under a specified null hypothesis. The associated p -value of the test is represented by the proportion of results at least as extreme as that of the original statistic from the initial sample. Permutation tests have the benefit that they require fewer assumptions than parametric tests (distributions do not have to be normal and samples do not have to be large); they tend to produce more accurate results than classical methods; they are widely generalizable (may be used for any statistic of interest, not just the mean or proportion); and they promote conceptual understanding. High school juniors and seniors seen as “less able” have been able to grasp the idea of permutation tests (Barbella, Denby, & Landwehr, 1990). Along with many other tools for mathematical and statistical investigation, *CPMP-Tools* contains a built-in feature specifically designed to conduct randomization tests in the context of curricular investigation. Hart, Hirsch, and Keller (2007) argue that the use of the randomization test in this technological environment may serve to amplify student learning of probability and statistics and “furnish a solid foundation for those students who go on to further study of statistical tests” (p. 199).

Students struggle with classical inferential statistics (Chance et al., 2004; Moore, 1997). Though resampling techniques will not likely replace the tradition of classical inference, they provide access for students to powerful, broadly useful, and conceptually-

based statistical ideas. It is conjectured that classical inference will be more accessible to students following experiences with resampling methods than without (Hesterberg, 2006). Though this conjecture was not tested as part of the research conducted here, it may be a productive line of research for the future.

Resampling methods are computer-intensive procedures requiring new forms of technology to manage massive data coordination. In order for teachers to begin to explore these ideas with students, they must learn to navigate in the technological environment required. With respect to teacher training, introduction to new technology and processes may “stimulate teachers to think about the processes of learning, whether through a fresh study of their own subject or a fresh perspective on students’ learning. It softens the barrier between what students do and what teachers do” (Bransford, Brown, & Cocking, 2000, p. 226).

Influence of Prior Research on the Present Study

Similar to the work of McClain, Cobb, and Gravemeijer (2000), a goal of the present study was that students (in this case teachers) will develop deep understandings of important statistical ideas (*comparing distributions*) as they use dynamic statistics software (*CPMP-Tools* and *Fathom2*) to represent data and make data-based arguments. Instead of focusing on computing various statistics and constructing graphs with little understanding, the environment that situates the present study supports the detective work of exploratory data analysis (EDA) while laying the groundwork for inference. Although the literature points to potentially productive courses of action to support student development of concepts such as distribution and variability, the literature fails to point to

productive courses of action for supporting teachers who have likely had the opportunity to study some statistics during their academic or professional life. It is possible and even likely that some of the same strategies for supporting students' statistical understanding may similarly support teachers.

Through utilizing the simulation features of *Fathom2* and *CPMP-Tools* while investigating data-intensive, context-rich situations in a professional development environment supportive of inquiry and collaboration, it is conjectured that teachers may develop improved conceptual understanding of *comparing distributions* and related statistical concepts. Of particular research interest were the ways in which simulation and resampling methods with *CPMP-Tools* and *Fathom2* shaped teachers' understanding of *comparing distributions* and distributions, variability, sampling distributions, and informal inference generally.

Design Considerations for Supporting High School Mathematics Teachers' Statistical Learning through Professional Development

This section attends to aspects of the study beyond research related to individuals' understanding of statistical concepts. In particular, literature regarding the importance of professional development for supporting teachers' understanding of statistical concepts, theoretical perspectives underpinning the study, and design research methodology are reviewed. The chapter concludes with a discussion of the conceptualization of the statistical content that would serve to inform and direct the construction of the hypothetical learning trajectory and assessment framework upon which this study is based.

*The Importance of Professional Development for Supporting
Teachers' Understanding of Statistical Concepts*

“By and large, teachers have a strong command of the procedural knowledge of mathematics, but they lack a conceptual understanding of ideas that underpin the procedures” (Mewborn, 2003, p. 47). Furthermore, Mewborn (2003) offers that professional development of teachers should put teachers’ thinking at the center of professional development and provide mathematical experiences in which opportunities exist for teachers to strengthen their conceptual understanding of topics as well as to make connections between topics. In the context of teaching and preparing teachers to teach statistics, it is useful to consider a construct similar to mathematical content knowledge for teaching (Ball, 2000, 2002, 2003; Ball & Bass, 2000) for statistics. It is not sufficient for teachers to simply know theoretical statistics. Instead, they “need mathematical (statistical) knowledge in ways that equip them to navigate . . . complex mathematical (statistical) transactions flexibly and sensitively with diverse students in real lessons” (Ball & Bass, 2000, p. 94). This flexibility applies to content, pedagogy, and use of technology in the context of statistics instruction.

Because of the complex nature of the interrelated big statistical ideas seen previously in this chapter and the lack of statistical preparation of most secondary mathematics teachers, the design, implementation, and analysis of the efficacy of professional development with the intent of supporting teachers’ statistical knowledge for teaching, was envisioned as important to the mathematics and statistics education research fields and potentially beneficial to the teachers who were participants in the study.

Based upon extensive research with the Local Systemic Change Initiative (LSC), Weiss (2006) considered the following to be features of high quality professional development: (1) the focus is on content knowledge, (2) active learning is emphasized, (3) coherence is promoted, (4) training is extensive and sustained over time, and (5) collaboration among teachers is encouraged. According to Loucks-Horsley, Stiles, and Hewson (1996), professional development for mathematics and science teachers should be designed according to seven principles: (1) Professional development experiences are driven by a clear, well-defined image of effective classroom learning and teaching; (2) they provide teachers with opportunities to develop knowledge and skills and broaden their teaching approaches, so they can create better learning opportunities for students; (3) they use instructional methods to promote learning for adults which mirror the methods to be used with students; (4) they build or strengthen the learning community of science and mathematics teachers; (5) they prepare and support teachers to serve in leadership roles if they are inclined to do so; (6) they consciously provide links to other parts of the educational system; and (7) they include continuous assessment. As will be seen throughout the remainder of this document, relevant characteristics associated with high quality professional development for teachers were intentionally designed into the professional development intervention.

Theoretical Perspectives

A situated perspective on knowing and learning with respect to teacher learning and professional development was adopted for the present study (P. Cobb, 2000; P. Cobb & Yackel, 1996; Greeno, 2003). As such, the focus was on how teachers became “more

successful in participating in statistical practices and how they develop[ed] identities as mathematical knowers and learners” (Greeno, 2003, p. 315). This perspective reflects the emergence of theories that view meaning, thinking, and reasoning as products of social activity (Lerman, 2000).

“The situative perspective . . . focuses researchers’ attention on how various settings for teachers’ learning give rise to different kinds of knowing” (Putnam & Borko, 2000, p. 6). Putnam and Borko (2000) suggested that although teachers’ learning may be intertwined with their ongoing practice, professional development outside of the classroom may stimulate teachers to consider knowledge in new and powerful ways. Because teachers’ patterns of thought and action may have become automatic in the classroom setting, the same patterns may be resistant to reflection or change. “Engaging in learning experiences away from this setting [the classroom] may be necessary to help teachers ‘break set’—to experience things in new ways” (p. 6). They suggest that summer workshops are particularly powerful settings for teachers to develop new relationships with mathematics and statistics content as well as insights into students’ learning. It is conjectured that teachers’ immersion into statistical investigations with unfamiliar technological tools and techniques may provoke sufficient disequilibrium from which teachers might begin to challenge their ideas about what it means to learn and do mathematics and statistics.

In the tradition of situated perspective, a multifocal lens for analysis is used. By coordinating analyses of individuals’ knowledge with the establishment and maintenance of group norms, trust, and collaborative interactions during group sessions, this research contributes to the knowledge base relating to how successful professional development

with high school teachers may be enacted. The “ways of knowing” typology (see below) provided a useful framework from which to consider teachers’ understanding of knowing and learning and it informed the creation of professional development materials. Informed largely by the work of Belenky, Clinchy, Goldberger, and Tarule (1986), Boaler and Greeno (2000) adapted the previous work to the mathematics world and suggested that “when mathematics learning practices place students in positions with more significant conceptual agency, it is easier for many of them to author their identities as learners with that kind of agency” (p. 196). Assuming that teachers in the present study played the role of student, this typology is useful.

- *Received knowing*, in which the individual considers her knowledge is primarily dependent on and derivative from an authoritative source other than herself.
- *Subjective knowing*, in which the individual considers her knowledge as primarily a result of her affective reactions to information and ideas.
- *Separate knowing*, in which the individual considers her knowledge as primarily being constructed to comply with rules that establish validity and to be defensible against challenges based on rules for validating knowledge.
- *Connected knowing*, in which the individual considers her knowledge as primarily being constructed in interaction with other people (either directly, in conversation, through interacting with texts or other representations of others knowledge and thinking), in a process that depends on understanding others’ experiences, perspectives, and reasoning, and incorporates this understanding into the individual’s knowing and understanding. (p. 196)

Boaler and Greeno (2000) speculated that teachers of mathematics tend to prefer received knowing, based upon their personal experiences with mathematics, and thus perpetuate the cycle of received knowers by teaching received knowing. The intent of the design of the professional development intervention in the present study was to support a more desirable connected knowing perspective among participating teachers that they may ultimately do so with their students.

Supporting productive disciplinary engagement of the high school mathematics teachers was another motivating force behind the design of the professional development intervention. Based on the work of Engle and Conant (2002), “productive disciplinary engagement can be fostered by designing learning environments that support (a) problematizing subject matter, (b) giving students authority to address such problems, (c) holding students accountable to others and to shared disciplinary norms, and (d) providing students with relevant resources” (p. 399). Though their work was in the context of elementary science students, it was hypothesized that the criteria for productive disciplinary engagement for fifth graders might not be all that different from that of all learners.

Three principles for learning mathematics that support and augment the recommendations of Engle and Conant (2002) are (1) engaging prior understandings, (2) the essential role of factual knowledge and conceptual frameworks in understanding, and (3) the importance of self-monitoring (National Research Council, 2005). The National Research Council (NRC) further elaborated design characteristics in support of effective classroom environments which impacted the design of this study’s professional development intervention. In particular, the NRC suggests teaching and learning environments should be learner-centered, knowledge-centered, assessment-centered, and community-centered.

Like Rubin and Hammerman (2006), a focus during the professional development in the present study was on “providing an environment in which teachers could explore important ideas about data and statistics using new software tools, and on conducting research on their thinking” (p. 19). Also, as Rubin and Hammerman elaborated, this study

assumed “learning is a slow, non-linear process of constructing and building more and more robust understandings over time” (p. 20). The study highlights the need for teachers to understand the complexity of arguments that may emerge when complex datasets are analyzed using dynamic technology with fairly effortless ability to create or modify representations. Since Rubin and Hammerman conducted a teaching experiment with a small group of teachers over the course of two years, they wondered what might happen with teachers in a shorter course. With the use of *Tinkerplots* for analyzing data, teachers rarely used measures of center to describe or compare graphs but chose to use different graphical representations, in particular, the binning features of *Tinkerplots*. Their study also raises the issue that when using tools like *Tinkerplots* or *Fathom2*, teaching requires “an in-depth understanding of the kinds of thinking the tool might engender and make visible. Once thinking is made visible, it can be discussed, challenged, and made more robust” (p. 37). This study investigated a shorter duration experience for teachers in order to begin to understand whether and in what ways professional development for high school mathematics teachers may effectively impact teachers’ statistical understanding.

*Design Research as an Appropriate Methodology in Statistics
Professional Development Research*

Based upon the paucity of research surrounding statistical professional development of high school teachers and the complexities identified in the literature related to the development of statistical thinking and reasoning, especially in an environment with new and flexible technology, this study utilized methods of design research. A design experiment was appropriate because it entails “both ‘engineering’ particular forms of learning and systematically studying those forms of learning within

the context defined by the means of supporting them” (P. Cobb, Confrey et al., 2003, p. 9). Design research commonly incorporates development of instructional materials, teaching experiments, and retrospective analyses (Bakker & Gravemeijer, 2004). Design experiments are structured subject to test and revision of theories through successive iterations.

According to Cobb et al. (2003), (1) the purpose of design experimentation is to develop a class of theories about both the process of learning and the means that are designed to support that learning; (2) design experiments are typically test-beds for innovation; (3) design experiments are prospective and reflective as they create conditions for developing theories yet must place these theories in harm’s way; (4) design is iterative as conjectures are tested and refuted, cycles of invention and revision dominate the process; (5) theories developed during the process of experiment are humble not merely in the sense that they are concerned with domain-specific learning processes, but also because they are accountable to the activity of design. These features are broadly applicable to design experiments, regardless of content area.

As design research aims to create and investigate innovative learning situations for purposes of understanding certain aspects of the learning situations and the ways in which learning may be supported, it is especially appropriate for research in statistics education (Gravemeijer & Bakker, 2006, p. 1). Cobb and McClain (2004) explicated design principles for the development of statistical instructional materials for students. These principles may have similar merit for the design of professional development for teachers.

These principles involve formulating and testing conjectures about:

- 1) Central statistical ideas, such as distribution, that can serve to orient the development of an instructional design
- 2) The characteristics of instructional activities that
 - a) Make it possible for students' classroom activity to be imbued with the investigative spirit of data analysis
 - b) Enable teachers to achieve their instructional agendas by building on the range of data-based arguments that students produce
- 3) Classroom activity structures that support the development of students' reasoning about data generation as well as data analysis
- 4) The characteristics of data analysis tools that
 - a) Fit with students' reasoning when they are first introduced in an instructional sequence
 - b) Serve as a primary means of supporting students' development of increasingly sophisticated forms of statistical reasoning
- 5) The characteristics of classroom discourse in which
 - a) Statistical arguments explain why the way in which the data have been organized gives rise to insights into the phenomenon under investigation
 - b) Students engage in sustained exchanges that focus on significant statistical ideas (p. 392)

The design of the professional development activity system in the present study incorporated the design principles recommended by Cobb and McClain (2004). The central idea that served to orient the study was *comparing distributions*. The statistical tasks were designed to promote the investigative experience as described as EDA which required teachers to make data-based arguments and to showcase their thinking and reasoning as the larger group worked to develop a shared understanding of the big ideas related to *comparing distributions*. The data analysis tools progressed from the use of graphing calculators, with which teachers were generally familiar, to the use of *CPMP-Tools*, statistical tools which may be seen as somewhat more route-based and less flexible, and then to use of *Fathom2* software for its flexible investigation potential. The classroom discourse and norms (both social and sociomathematical) were negotiated with an eye toward productive discourse and argumentation focused on significant aspects of

comparing distributions and related big ideas. The design and implementation of the professional development intervention is the focus of Chapter IV.

*Toward a Basis for Statistical Content Inclusion in Professional Development
for High School Mathematics Teachers*

Based upon a review of the literature related to students' and teachers' understanding of *comparing distributions*, distributions, variability, sampling distributions, and informal inference, the overlap among ideas required careful consideration of structuring concepts that would be productive for the design of learning environments for teachers. Drawing from the literature reviewed in this chapter, a table of the big statistical ideas related to *comparing distributions* and supporting sub-ideas potentially important to consider in the context of professional development for high school teachers was proposed (Table 4). Bakker and Frederickson (2005) recognized variation, sampling, data, and distribution as key concepts for students at the middle school level which could be supported through *comparing distributions* and growing samples. The present study assumes a similar position but also incorporates sampling distributions and informal inference, which are suggested for study by high school students in PSSM (NCTM, 2000), GAISE (2006), and seen as recurring themes in the literature. Table 4 served as a design tool during the construction and sequencing of learning activities in the professional development intervention.

Figure 2 captures the essence of the ways in which these big ideas and related sub-ideas may be conceptually connected. *Comparing distributions* was broadly conceived to encompass a triadic, multidirectional relationship between the statistical ideas of distribution, variability, and sampling distributions, all potentially supporting

informal inference. This conceptual framework was used to guide the design of the professional development intervention, assessment instruments, and quantitative analysis.

Table 4

Big Ideas, Sub-Ideas, and Considerations Related to “Comparing Distributions”

Distribution	Variability	Sampling Distributions	Comparing Distributions	Informal Inference
- Center - Typical - Average - Algorithmic - Balance Point - Median - Data Reducer - Mode - Signal and noise - Spread - Types of variability - Shape - Density - Skewness - Data as pointer - individual cases - classifier - as an aggregate - Representations - Dot plots - Stem plots - Histograms - Box plots - Context - Growing samples - Student-generated Data - Representations - Informal language - Predictions	- Variability due to errors of measurement - Variability due to multiple causes - Sample to sample variability - IQR - Standard deviation - MAD (mean & median) - Particular values—extremes/outliers - Change over time - Range - Likely range - Distance to some fixed point - Sum of residuals - Covariation or association - Distribution - Noticing & acknowledging - Measuring and modeling for the purposes of prediction, explanation, or control - Explaining and dealing with - Investigative strategies	- Distributions - Variability - Compare sample to population - Physically draw samples - Technology exploration - Discuss student observations - Assess understanding - Build on understanding - Growing samples - Sampling experiment - Make and test predictions - Simulations - Resampling - Randomization test	- Numerical strategies - Visual strategies - Experiments - Measurable conjectures - Tolerance for variability - Understanding context - A view towards inference	- Reasoning from distributions - Reasoning w/centers - Boxplots w/dotplots - Properties of aggregates - Sample size - Controlling for bias - Tendency - Sampling variation - Null hypotheses - p-values - Significant difference

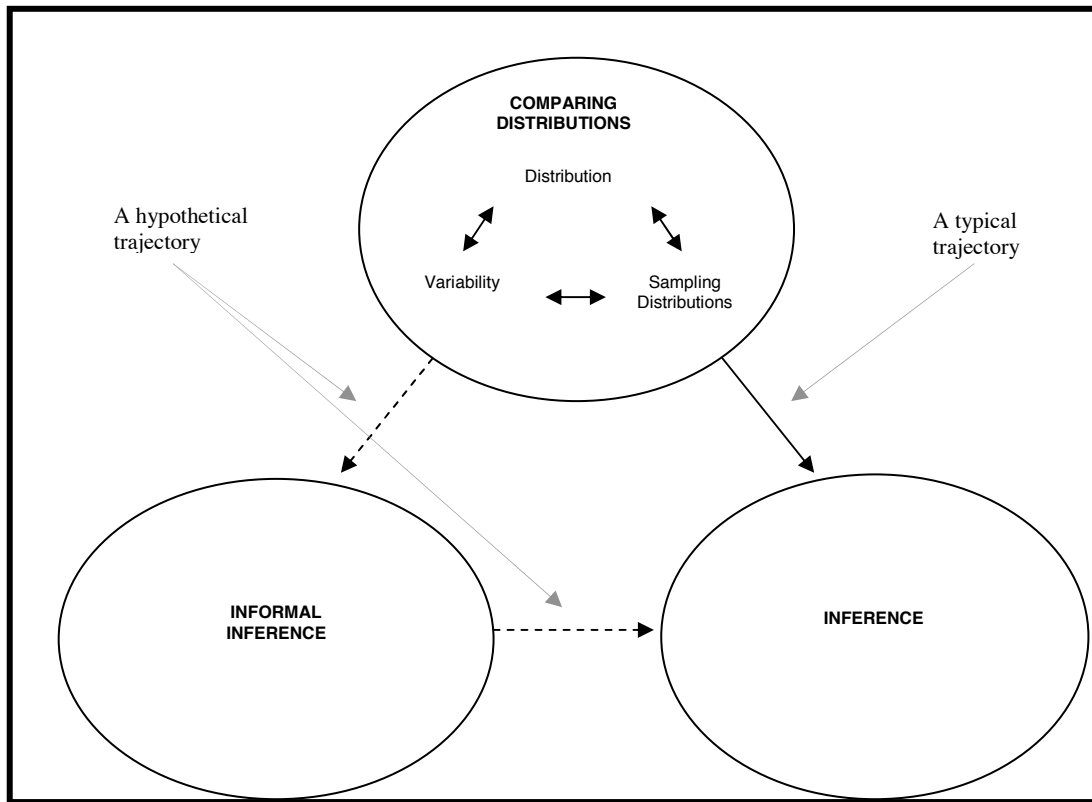


Figure 2. A conjectured conceptual relationship between big statistical ideas.

Building on prior research documented in this chapter, methodology of design experiments, and the conceptual and theoretical frameworks discussed herein, Chapter IV presents the initial hypothetical learning trajectory, the emergent learning trajectory, design and assessment considerations, and detailed episodes from video documentation of the professional development intervention.

CHAPTER III

METHODOLOGY

Theoretical Framework

This study was conceptualized as a design experiment as it entailed “both ‘engineering’ particular forms of learning and systematically studying those forms of learning within the context defined by the means of supporting them” (Brown, 1992; P. Cobb, Confrey et al., 2003, p. 9). As reported in Chapter II, little is known about high school teachers’ understanding of big ideas of statistics, other than the speculation that it is not very strong (CBMS, 2001; Garfield & Ben-Zvi, 2004; Makar & Confrey, 2004). Given that statistics is now consistently recommended to be an important area of study for all students (The College Board, 2006; Franklin et al., 2007; NCTM, 2000, a study of teachers’ understanding of *comparing distributions* is both important and timely. Furthermore, the field is still wrestling with what constitutes effective professional development (Loucks-Horsley, Love, Stiles, Mundry, & Hewson, 2003; Weiss, 2006). The design experiment in the context of mathematics professional development for high school teachers provides a methodology with potential for helping to illuminate or elaborate possible theories and mechanisms that may support future work in this area.

Understanding of *comparing distributions* is broadly construed to incorporate understanding of “distributions,” “variability,” and “sampling distributions,” and the relationships among them. In order to characterize teachers’ understanding of *comparing*

distributions, the study draws heavily on several frameworks developed by other researchers to characterize students' understanding of *comparing distributions* (Makar & Confrey, 2002), variability (Watson et al., 2003), and sampling distributions (Chance et al., 2004). Though Makar and Confrey's (2002) framework was specifically designed for representing levels of reasoning of *comparing distributions*, it was insufficiently detailed for the purposes of the current study (see Table 5). Because of the conceptualization of *comparing distributions* in the present study, the additional sampling distribution and variability frameworks were essential for discriminating between levels of understanding of *comparing distributions*. Through the coordination of these three frameworks and their impact on the development of question-specific rubrics, pre- and post-assessments of teachers' understanding of *comparing distributions* could be meaningfully evaluated and compared.

Table 5

Coordinating Multiple Frameworks—Comparing Distributions, Variability, and Sampling Distributions

Comparing Distributions	Level 0 Blank	Level 1	Level 2	Level 3		Level 4
Makar & Confrey (2002) Comparing Distributions	Level 1—Pre-descriptive Level	Level 2—Descriptive Level	Level 3—Emerging Distributional Level	Level 4—Transitional View		Level 5—Emerging Statistical Level
Watson et al. (2003) Variability	N/A	Level 1—Prerequisites for variation	Level 2—Partial recognition of variation	Level 3—Applications of variation		Level 4—Critical aspects of variation
Chance, delMas, & Garfield (2004) Sampling Distributions	N/A	Level 1—Idiosyncratic Reasoning	Level 2—Verbal Reasoning	Level 3—Transitional Reasoning	Level 4—Procedural Reasoning	Level 5—Integrated Process Reasoning

Given that a paucity of research exists regarding effective ways for high school teachers to learn important statistical content knowledge for teaching (Makar & Confrey, 2004), this study was envisioned as what Borko (2004) defined as Phase I: Existence Proofs of Effective Professional Development. During Phase I, research is typically conducted at a single site with the intention of exploring whether and in what ways teachers' learning can be supported. Borko notes that in these instances, the professional development designers are usually also the researchers. She suggests the situative perspective is appropriate in order to attend to individuals as a unit of analysis, the group as a unit of analysis, and appropriate frameworks and tools that help to coordinate the two perspectives simultaneously. This is consistent with what Cobb and Yackel (1996) refer to as the emergent perspective.

A situated perspective on knowing and learning with respect to teacher learning and professional development was adopted for this study (P. Cobb, 2000; P. Cobb & Yackel, 1996; Greeno, 2003). As such, the focus was on whether and how teachers became more successful in participating in statistical practices and how they developed identities as mathematical knowers and learners (Greeno, 2003). This perspective reflects the emergence of theories that view meaning, thinking, and reasoning as products of social activity (Lerman, 2000).

It was hypothesized that this research may provide an existence proof that high school mathematics teachers' knowledge of *comparing distributions* could improve significantly, given a well-designed and implemented professional development experience. From the design perspective, it was hypothesized that an environment in which teachers explored statistical ideas from a problem-based perspective with flexible

and dynamic technological support and in which norms of participation included a strong emphasis on reasoning and justification, had potential for positively impacting teachers' understanding of *comparing distributions*. Additionally, a working hypothesis was that an effective professional development experience, in the area of statistics, might propel teachers to reconsider what it means to learn and do mathematics and statistics in a problem-based environment with access to multiple modes of technology. Underlying these hypotheses was the belief that for teachers to promote robust student understanding of big mathematical or statistical ideas, they must be mathematically and statistically proficient with knowledge that is both rich and flexible (Borko, 2004; National Research Council, 2001). The professional development intervention for this study was designed and implemented with these goals in mind. Details of the development and enactment of the intervention may be found in Chapter IV.

Because the researcher, professional development designer, and professional development facilitator is the same person in this study, a knowledgeable skeptic may question the trustworthiness and generality of the study (Schoenfeld, 2007). In response to these potential concerns, the researcher's major advisor reviewed and commented on the design of the professional development materials. Furthermore, to provide another researcher's perspective, a second researcher, a university mathematics educator, assisted the primary researcher in several important ways: (1) reviewed and commented on the design of the professional development materials, (2) attended and video-taped all professional development sessions, (3) debriefed and problem-solved with the researcher during and after each day's session, (4) scored a large sample of open-ended items on the

pre- and post-assessment to establish inter-rater reliability, and (5) responded to written portions of this work in order to confirm and/or refute the validity of the claims.

The next sections will describe the sites for the study, the participants, the general conduct of the study with associated timeline, and the analyses used in the study.

Research Sites

Fifty-six (56) high school mathematics teachers from 22 school districts participating in a state-level Mathematics and Science Partnership (MSP) project were the participants for this study. The MSP was the umbrella project within which this study was situated. All school districts with teachers in this study were characterized as rural districts. For professional development, the 56 participating teachers were distributed across three distinct sites within one state in the Midwest. Teachers were not randomly assigned to sites, rather each region was determined by geography and teachers' abilities to attend professional development in a particular geographic location during a specific timeframe. Table 6 reflects the composition of the three professional development sites by school and number of participating teachers. Demographic and student performance data are included to provide a view of the landscape which situates the professional work of teachers in this study. Most of the schools are quite small with a moderate to large percentage of students eligible for free or reduced-price lunch and with only 6 of 22 schools reporting 60% or more 11th-grade students proficient in mathematics. In some cases teachers from the same district attended different regional sessions. Each of the regions consisted of teachers from relatively small schools ($n < 500$ students), medium-sized schools ($500 < n < 1000$ students), and large schools ($n > 1000$ students).

Table 6

Regional Site Distribution and 2006-2007 Demographic Data for Schools of Participating Teachers

2006-2007 Data	# of participating teachers	# of grades 9-12 students 2005-2006 NAEP (nces.ed.gov)	% of free/reduced lunch eligible students 2005-2006 NAEP (nces.ed.gov)	% of grade 11 students considered proficient in mathematics (state assessment) Class of 2006
Site 1				
School A	1	228	88	43
School B	6	775	39	60
School C	3	263	23	76
School D	6	1116	26	56
School E	3	395	32	59
School F	1	126	23	61
School G	2	388	29	60
School H	2	390	Not available	54
School I	1	71	55	40
School J	2	232	70	50
School K	1	95	59	48
School L	1	238	48	58
School M	2	362	28	48
13 schools	31 teachers			
Site 2				
School N	3	267	54	43
School O	1	597	39	46
School P	2	400	67	45
School H	1	390	0?	54
School Q	1	528	59	32
School D	1	1116	26	56
6 schools	9 teachers			
Site 3				
School L	1	238	48	58
School R	7	1656	30	66
School N	2	267	54	43
School S	1	108	80	38
School T	2	341	73	76
School U	2	187	48	58
School V	1	572	55	43
7 schools	16 teachers			
			Average free/reduced for group 42% (excluded School H)	State average 52.4%

The MSP was a two-year professional development research partnership between a Midwestern university, three state-level Mathematics and Science Centers, and 23 rural school districts. One high school sent no teachers to the professional development

program. The primary goals of the project were (1) to improve teachers' mathematics content knowledge for teaching, (2) improve articulation between middle and high school mathematics content and instruction, and (3) ultimately to improve student achievement as measured by the state-administered assessment. The MSP began July 1, 2005 and continued through August 31, 2007.

The MSP project provided extensive content- and pedagogy-focused professional development to teams of middle school and high school teachers of mathematics through school year professional development, summer institutes, and district-based professional learning communities. One component of the MSP was an 8-day professional development summer institute. The final four days of the summer institute, in each of the three geographical locations, was the setting for the present study and involved only high school teachers. Prior to the professional development program being investigated in this study, the middle school and high school mathematics teachers had participated in three full-day professional development sessions (October 2005, December 2005, and February 2006), and four consecutive full-day sessions during the first half of the summer institute (June 2006). The mathematical content foci of the professional development during those sessions were algebraic and geometric reasoning. Approximately half of the teachers had participated in four three-hour district-level professional learning community (PLC) sessions as part of an experimental treatment associated with the MSP.

The Setting

The professional development experience took place over four consecutive days, (approximately 20 hours) for each of the three groups of high school teachers described

previously. The professional development was designed to incorporate the use of resampling techniques and dynamic statistical software (*CPMP-Tools* and *Fathom2*) to investigate situations involving the comparison of two or more distributions. The professional development was largely investigative in nature, using a variety of tasks from physical simulations to computer simulations in order to model real-world phenomena and support teachers' thinking and reasoning with respect to *comparing distributions*. Consistent with the theoretical framework, the investigations and associated conversations and reflections also supported the development of the big ideas of distribution, variability, and sampling distribution, all supporting ideas of informal inference. The design of the professional development materials is elaborated in Chapter IV.

The data collection schedule for the study is shown in Table 7.

Table 7

Data Collection Schedule for Summer 2006

June	July	August			
Content/ Background Pretest <i>n</i> = 56	Pre-Intervention Interviews <i>n</i> = 9		M-T-W-Th		F
		Week1	PD Site 1 <i>n</i> = 31	Post-test Site 1	
		Week 2			Post-Interview Site 1
		Week 3	PD Site 2 <i>n</i> = 9	Post-test Site 2	
		Week 4	PD Site 3 <i>n</i> = 16	Post-test Site 3	Post-Interview Site 2
		Week 5			Post-Interview Site 3

Participants

Mathematics teaching experience for teachers ($n = 56$) in this study ranged from 1 to 40 years, with mean 10.1 years and median 9.5 years. Figure 3 presents a histogram of the distribution of years of teaching experience.

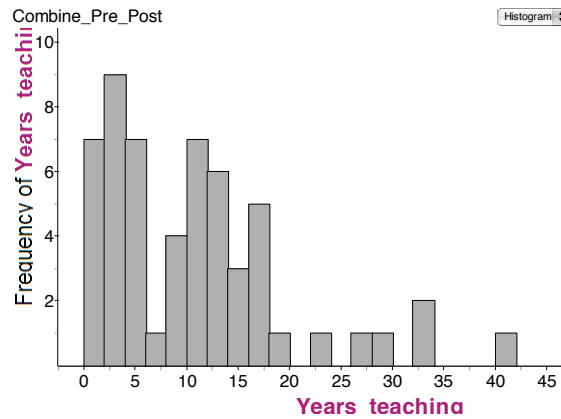


Figure 3. Distribution of participating teachers' number of years teaching mathematics.

The group might be characterized as moderately experienced; however, 23 of 56 (41%) of the teachers were within their first five years of teaching. All of the teachers in this study met the No Child Left Behind (NCLB) requirement of “highly qualified” status and all but one teacher was secondary-certified to teach mathematics. Self-reported certification status is provided in Table 8.

Table 8

Participating Teachers' Reported Certification

Certification	Yes	No
Regular Education	45	11
Secondary	55	1
Elementary	1	55
Math Major	42	14
Math Minor	14	42

In a survey administered as part of the MSP within which this study is situated, 23 of 56 (41%) teachers reported no teaching of data analysis and statistics in any of their mathematics courses during the school year directly preceding this project. An additional nine (16%) teachers reported teaching some data analysis and statistics in some classes, but not in the target class about which they were responding. Of the 33 teachers reportedly teaching data analysis and statistics at some time during the year, 25 (76%) rarely or never introduced software to help develop and support students' statistical reasoning. Collectively, these data suggest that data analysis and statistics may not have been strongly integrated throughout mathematics programs of this group of teachers.

A component of the statistical content pre-assessment (see Appendix A) administered prior to the professional development associated with this study (June 2006), contained a number of prompts designed to provide some indication of teachers' statistical background and comfort-level (Makar, 2004). Indicators ranged from 1 to 5 with 1 representing low comfort-level, 5 representing high comfort-level. As can be seen in Table 9, teachers reported highest comfort-levels for statistical graphs, followed by descriptive statistics. The remaining categories were rated substantially lower by teachers, with sampling distributions and statistical inference scored 1 by 75% or more teachers. Given that the majority of these teachers are secondary-certified and that 30 (54%) reported taking one statistics course, 14 (25%) reported taking two statistics courses, and 2 (4%) reported taking three or more courses, the self-reported data suggest that the knowledge residue from the courses taken by these teachers may not have been terribly strong or, perhaps these teachers did not feel comfortable with many big ideas from statistics as a result of their participation in university courses. Seven (12%)

teachers reported never having a statistics course and 3 (5%) teachers left the question blank.

Table 9

Teachers' Self-Reported Comfort Level for Big Ideas from Statistics

	Level of Comfort ($n = 56$)					
	1	2	3	4	5	M
Descriptive statistics (mean, standard deviation, z -score)	1	17	19	14	5	3.09
Statistical Graphs (histogram, boxplot, bar graph)	0	3	11	23	19	4.04
Distributions (normal, chi-square, probability density functions)	28	15	10	2	1	1.80
Experimental Design (surveys, blocking, bias, sampling methods)	21	21	9	4	1	1.98
Correlation and Regression (least squares, r^2 , residuals, outliers)	21	22	10	1	2	1.95
Sampling Distributions (Central Limit Theorem)	42	11	1	0	2	1.37
Statistical Inference (t -tests, confidence intervals, chi-square tests, power, Type II error, ANOVA)	45	6	2	2	0	1.29 (1 blank, $n=55$)

As an aggregate indicator of teachers' comfort-level with these statistical ideas, an indicator was created by averaging each teacher's ratings across all seven categories.

Figure 4 displays the distribution of the average comfort-level indicator for all 56 teachers in this study ($M = 1.60$, $Mdn = 1.65$, $SD = 0.43$). According to this indicator, teachers' aggregate comfort-level with statistics was quite low, with the exception of two teachers. Those two teachers, as well as other teachers with experience teaching statistics as either a stand-alone course or in an algebra course, tended to rate their comfort-level with statistical ideas slightly higher than those with no experience.

A subset of the 56 participants was selected to participate in pre- and post-professional development interviews.

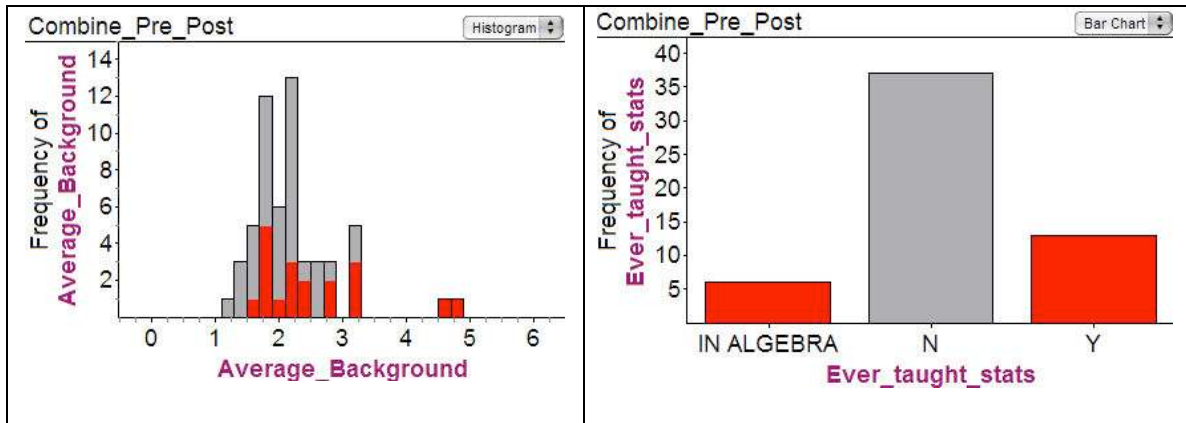


Figure 4. Teachers average comfort-level with statistical ideas compared to experience, low = 1, high = 5.

Interview Participants

In order to more fully understand and attempt to capture the lived experiences of individuals in this study, it was essential to conduct participant interviews. Initial interviews were conducted prior to the professional development but after the statistics content pre-assessment (July 2006). The pre-professional development interviews were conducted to provide additional insight, beyond that from the written assessment, into teachers' thinking and reasoning about *comparing distributions* and their prior statistical experiences in order to (1) inform the design and development of the professional development program, and (2) to provide a richer characterization of teachers' present understanding than was possible in a written assessment. The interviews were semi-structured, lasting 30 to 45 minutes, and used the protocol found in Appendix C.

In order to provide representation and teachers' voices from each of the professional development sites, it was important to select interview candidates from each of the three geographic locations at which the professional development programs were conducted. The ultimate selection of participants was based upon their (1) pre-assessment

statistics content assessment performance, (2) availability during the time-frame required for this study, and (3) willingness to be interviewed for this study. The individual level of performance on statistics content assessment was used to select teachers representing a range of performance. Prior to beginning the study, it was determined that from each of the three regions, three people would be interviewed: one relatively low performer on the pre-assessment, one relatively medium performer on the pre-assessment, and one relatively high performer on the pre-assessment. In order to guarantee sufficient pre- and post- interview data to inform the study in case of participant attrition or change of venue from June to August, 20 pre-interviews were conducted (Site 1–7, Site 2–7, Site 3–6). All interviews were audio and videotaped.

The post-interviews were conducted one week following each four-day professional development program. Nine interviews were conducted at Site 1, 4 at Site 2, and 3 at Site 3, for a total of 16. Post-interviews were similar to pre-interviews except that the background questions were omitted and one question following up on assessment task 13 was omitted because its potential to inform the study was determined to be minimal based on the pre-interview responses. It was decided that teachers' performance on a problem-based task would have more potential for informing the study than would responses to task 13, thus the substitution was made on the post-interview to include a *comparing distributions* task requiring use of technology. The assessment instrument is discussed in a later section of this chapter. The post-interview protocol may be found in Appendix D. Post-interviews lasted between 45 and 80 minutes and were again audio and videotaped.

A decision to reduce the number of pre-post paired interviews for transcription and analysis from 16 to 9 was based on several factors: (1) the original plan was to conduct nine sets of interviews, (2) time and cost to transcribe and analyze interview data was extensive, (3) the researcher suggested that the potential to inform the study was not equal across interviews, (4) maintaining the balance of three sets of interviews for each region could be achieved. Once this decision was finalized, interviews to be transcribed for analysis were determined by (1) comparing pre- and post-assessment performance, (2) assessing potential to contribute to the study on the basis of willingness to speak openly during the interview, and (3) attention to representation by gender. Table 10 summarizes characteristics of the teachers interviewed whose contributions were analyzed in this study. These data illustrate the comparability between the interview sample and all 56 participating teachers with respect to statistics courses taken or taught and aggregate comfort-level with statistics. The interview sample has slightly less experience teaching mathematics than the larger group, but the number of statistics courses taken or taught and general comfort-level with statistics is comparable. All names have been replaced with pseudonyms. Analyses from the transcribed interviews are presented in Chapter V.

Data Collection/Instrumentation

The entire professional development program was videotaped for use during the retrospective analysis and theory building following the intervention. The researcher and the videographer additionally took field notes during the professional development programs and met daily to debrief, adjust conjectures, and modify the hypothetical learning trajectory as part of the iterative daily minicycles of ongoing continual improvement to the design (P. Cobb, 2000).

Table 10

Background Characteristics of Interview Subjects

Name	Region	Gender	# years teaching math	# stats courses taken	Aggregate comfort-level	Ever taught statistics?
Cameron	1	M	16	1	1.71	N
Jessy	1	F	8	1	2.29	N
Lorraine	1	F	2	1	1.57	N
Jaylee	2	F	1	2	1.42	N
Sasha	2	F	5	1	2.14	Y
Jordan	2	F	3	1	2.83	Y
June	3	F	10	1	1.57	In algebra
Callie	3	F	18	1	1.57	N
David	3	M	8	2	4.57	Y
Average, $n_{\text{interview}} = 9$			7.89	1.22	2.18	Y-3, N-6
Average, $n_{\text{all}} = 56$			10.13	1.18	2.19	Y-19, N-37

Pre- and Post-Assessments

Written pre- and post-assessments of statistical content knowledge were created and may be found in Appendices A and B. The written pre-assessment instrument contained two background information questions to provide a baseline of statistics history for each participant. The instruments listed seven statistical ideas for which teachers were asked to rate their comfort-level with each idea using a scale from 1 (low) to 5 (high). Additionally, the instrument consisted of 21 prompts, of which 12 were multiple-choice and 9 were open-ended, constructed-response tasks. The assessments were designed to capture teachers' thinking about *comparing distributions* while at the same time considering, characteristics of distributions, variability, and sampling distributions. Seventeen of the items and two of the scoring rubrics were adapted from the work of Makar (2004). Many of these items were adapted from the work of other researchers; therefore validity was increased as results may be compared across studies. Three additional tasks were written, reviewed by statistics and mathematics education faculty

members, and piloted for this study. Several changes were made to the tasks based upon the feedback received. As will be further explained in Chapters IV and V, tasks were scored either individually or in related clusters using task- or cluster-specific four-level scoring rubrics.

Pre- and Post-Interviews

Pre- and post-interviews were conducted with nine participating teachers, three from each of the professional development sites, in order to provide additional detail related to teachers' background with statistics, their understanding of statistical ideas, and their reaction to the professional development experience. The interviews were semi-structured, using the interview prompts found in Appendices C and D. Interviews were audio and videotaped for transcription and analysis.

Written Reflections

During the course of the professional development program, teachers were asked on five separate occasions to respond to written reflection prompts. The reflection prompts may be found in Appendix E. Teachers' reflections provided formative assessments for the instructor, influenced the instructional design, and contributed to the retrospective analysis at the end of the intervention.

Analysis of Data

In line with the emergent perspective (interpretive approach), data were analyzed from both psychological and sociological perspectives. In order to better handle the complexity of the learning environment, like Cobb (2000), the researcher recognized the

need for an analytical approach which helps to identify any signals within the spectra of noise found in a classroom setting. The analysis was used to (1) inform and improve the instructional design, (2) document the collective mathematical learning of the professional development community, and (3) document the developing statistical reasoning of individual teachers as they participated in the practices of the professional development community. During the ongoing professional development experience, what was created was a record of knowledge. All of the evidence was interrogated to inform theory building.

According to Goldin (2003),

one continuously makes inferences about the internal states of others, based on their production of, or interaction with, representations external to them. . . . mathematical power requires competence in standard representation and their manipulation. But it also includes the ability to recognize and visualize structural relationships; to think spatially; to generalize and particularize; to formulate problem-solving strategies; to employ a variety of heuristic techniques and creative methods; and to experience such feelings as curiosity, bewilderment, frustration, purposefulness, elation, and satisfaction as appropriate. Mathematical concepts are learned powerfully when a variety of appropriate internal representations, with appropriate relationships among them, have been developed. (pp. 277-278)

Because we cannot directly observe normative taken-as-shared meanings or individual teachers' meaning based in mathematical activity, the emergent perspective forces us to "develop and test conjectures about both communal mathematical activities (social perspective) and individual students' reasoning (psychological perspective) as we analyze what the teachers and individual students say and do in the classroom" (P. Cobb, 2000). In this way, the interpretation consists of coordinating two different views to try to better understand and make sense of the things that are going on in the classroom. In this case,

the classroom is that of the professional development site. Yackel and Cobb (1996) coordinated these perspectives as seen in Table 11.

Table 11

An Interpretive Framework for Analyzing Classroom Mathematical Activity and Learning

Social Perspective	Psychological Perspective
Classroom social norms	Beliefs about own role, others' roles, and the general nature of mathematical activity in school (professional development)
Sociomathematical norms	Mathematical beliefs and values
Classroom mathematical practices	Mathematical interpretation and reasoning

Because the participants in the study were geographically clustered into three groups, there were three distinct sets of high school teachers in this study. The multiple-group structure allowed for the professional development intervention to undergo two separate rounds of revision. Following the first four-day session at Site 1, in addition to the daily analysis and revision of plans, a relatively thorough reflection of the session as a whole was undertaken in order to prepare for the second four-day session. Between the second and third iteration, less time was available for analysis; however, plans and activities were modified slightly based upon the collective insights of the primary researcher and assistant. Daily debriefing sessions allowed the events of each day to be interpreted from two researchers' perspectives and used to plan and/or modify prospective investigations, while the retrospective analysis of the entire design experiment occurred after all three sessions were completed.

Answering the Research Questions

1. What do high school mathematics teachers know about *comparing distributions*?

2. How do professional development experiences with resampling techniques and dynamic statistical tools, as described in this study, shape what teachers know about *comparing distributions*?
3. What characteristics of professional development for high school mathematics teachers contribute to their understanding of *comparing distributions*?

To answer the research questions, analyses sensitive to teachers' content knowledge as well as change in content knowledge were needed. Of particular challenge was that of coordinating the multiple data sources and analyses to do two distinctly different, but related things. First, ongoing analysis was used to inform the development and amendment of the hypothetical learning trajectory that guided the professional development intervention. Second, retrospective analysis was used to answer the research questions. Ongoing analyses included attending to pre-assessments, pre-interviews, review of written teacher reflections, and debriefing reflections on the professional development experience with the express purpose of influencing the hypothetical learning trajectory. Retrospective analysis required attention to comparing pre- and post-assessments and interviews, analysis of teacher written reflections, and examination of video from the professional development programs to identify changes in teacher understanding and to conjecture as to the impetuses for any changes. Making the connection between the hypothetical learning trajectory and changes in teacher understanding of *comparing distributions* required both levels of analysis.

Answering Research Question 1

Analyses of teacher pre-assessments and transcripts of pre-interviews were used to answer the first research question, *What do high school mathematics teachers know about “comparing distributions”?* Twenty-one pre-assessment items were clustered to represent 10 big ideas related to *comparing distributions* in accordance with the literature review. Each cluster-item was then evaluated against a four-level rubric which had been developed by the primary researcher specifically for each cluster and in conjunction with the theoretical framework for the study. The rubrics may be found in Appendix F.

All data were entered into a database for analysis. To establish reliability in scoring, the following process was used: (1) the primary researcher scored each item; (2) a random sample of 25% of the pre- and post-assessments were then scored by a second researcher prior to any discussion of the intent of the rubrics (this decision was purposeful in order to determine the clarity of the rubrics); (3) the two sets of scores were compared and discrepancies were investigated in order to more clearly refine the rubrics, agreement was reached on all items; (4) the primary researcher then blindly rescored all pre- and posttest items using the refined rubrics and then compared the second scores to the initial scores; (5) when discrepancies were found, items were carefully analyzed and scores reconciled; and (6) finally, for each item, pre- and post-assessment responses were combined and sorted by score. Each score-level cluster was reviewed for consistency and as a final check for possible errors. Once items were scored, pre- and post-assessments were analyzed quantitatively. Aggregate scores for individuals were determined by computing the mean of the ten cluster scores, producing an overall measure with range

0-4. Similarly, sub-strand measures representing understanding of (1) distribution, (2) variability, and (3) sampling distributions were created and analyzed.

Interview transcripts from the subset of nine teachers were analyzed qualitatively using techniques discussed by Miles and Huberman (1994). Because of the content-nature of the analysis, a beginning set of codes was created, applied, and modified to the transcripts. Because of the conceptual framework underpinning the professional development intervention, codes associated with *comparing distributions*, distribution, variability, and sampling distributions and their various dimensions, represented a good starting point. This procedure represented a compromise between an entirely a priori coding approach and a completely inductive approach. Throughout the coding process, codes were created or amended as needed in order to give voice to the data. The analysis of the pre- and post-interviews augmented that of the content assessments.

Additionally, to explore changes in teacher understanding due to the professional development experience, the written pre- and post-assessments were analyzed quantitatively using a matched-pairs pre/post design. Descriptive statistics, *t*-test, ANOVA, and non-parametric procedures were used to examine change from pre- to post-assessment within and across each of the three professional development regions and to determine whether recorded differences were statistically significant. For all quantitative comparisons in this study, a maximum alpha level of 0.05 was used to determine statistical significance. Subsets of the assessment items were analyzed separately to reflect emphases on understanding distributions, variability, and sampling distributions. Evidence of significant change was used to inform the continued qualitative analysis of interview data and video of the professional development.

Answering Research Question 2

The second research question, *How do professional development experiences with resampling techniques and dynamic statistical tools, as described in this study, shape what teachers know about “comparing distributions”?*, was answered through a combination of analyses. Although items on the content test did not involve the use of dynamic software by teachers, several items presented teachers with graphical and numerical representations from which to reason which would have been similar to those they might otherwise produce with dynamic statistical tools. Thus, a first cut included looking at whether and how teachers’ responses to these items changed from pre- to post-assessment. Interview transcripts were examined for references to resampling and dynamic statistical tools. The post-interview protocol included a task requiring teachers to use a dynamic statistical tool of their choice to conduct a brief statistical investigation; data from the interviews provided specific evidence of teachers’ use of dynamic statistical tools. Professional development video was reviewed to look for evidence of whether and how use of resampling techniques and dynamic statistical tools might be shaping teachers’ understanding of *comparing distributions*. An important component of answering this question involved the analysis of the twenty-four written reflection responses collected from the teachers throughout the professional development experience.

One particular challenge was to use the multiple sources of data to coordinate and triangulate findings, while also providing counterexamples to developing theory. Grounded theory (Strauss & Corbin, 1998) was used to conduct the retrospective analysis of the corpus of data in order that theory would emerge from the data. Because grounded

theory is responsive to the situation in which the research is done, one is continually looking for evidence to disconfirm the emerging theory. Ultimately, the emerging theory is considered adequate if it fits the situation and if it works. In this case, grounded theory helped to make sense of teachers' learning in a professional development setting, with a focus on *comparing distributions* and the use of dynamic statistical tools and resampling techniques, and to inform future work in this area. In line with this methodology, the literature review for this study continued to emerge as the study unfolded.

Answering Research Question 3

The third research question, *What characteristics of professional development for high school mathematics teachers contribute to their understanding of 'comparing distributions'?*, was answered through retrospective analysis. The analysis incorporated review of the videotapes and field notes taken during the professional development program intervention and analysis of the written teacher reflections. The pre- and post-content assessments were examined for their potential to provide initial direction for the analysis. In particular, items or combinations of items for which the greatest pre/post changes were evident, served as catalysts for reviewing the video, field notes, and reflections. The changes in assessment items helped to situate the analysis in the context of the professional development intervention and provide a backdrop against which to begin to conjecture as to the possible rationale for the changes in assessment scores. The task involved in answering the third research question was to interrogate the entire corpus of data to hypothesize about what aspects of the professional development experience

may have provoked changes in assessment scores of teachers. As with question 2, a grounded theory approach was used for this analysis.

Organization of Remaining Chapters

The reader will likely sense the difficulty of the presentation of the results of a study like this. A completely linear presentation of the study and related results is impossible and the reader will be referred to chapters and appendices that may need to be referenced along the way. The tensions between the emerging understanding of statistical content, the use of technology, the participants, and the researcher represent complex dynamics to capture in prose. Because several data sources were commonly used for ongoing and retrospective analyses, there is some duplication in presentation. Chapter IV incorporates analyses that impacted the design, evolution, and implementation of the hypothetical learning trajectory. Chapter V presents analyses with primary focus on the research questions. Finally, Chapter VI synthesizes the results from Chapters IV and V in order to answer the research questions, present a summary of the findings for this study, and address implications of this research, limitations, and potential directions for future study.

CHAPTER IV

DESIGN AND IMPLEMENTATION OF THE PROFESSIONAL DEVELOPMENT INTERVENTION

This chapter provides details and background of the design and modification of the four-day, professional development hypothetical learning trajectory (HLT) investigated in this study. The preliminary HLT is presented and followed by discussion of guiding principles for the design of the professional development intervention. The rationale for the selection and structuring of learning activities is discussed. Next, the alignment between the HLT, the big ideas associated with *comparing distributions*—distribution, variability, and sampling distributions, and the assessment instrument are presented. Finally, the enacted HLT is elaborated and justifications for the modifications to the initial HLT are discussed.

What Is a Hypothetical Learning Trajectory?

A hypothetical learning trajectory is defined as “the learning goal, the learning activities, and the thinking and learning in which students might engage” (M. A. Simon, 1995, p. 133). It is expected that the hypothetical learning trajectory will evolve as the facilitator and students co-construct the environment in which the intentioned learning is taking place. As the facilitator becomes more informed of students’ learning through analysis of the actions of the students and reflection on the activities in the classroom, the trajectory may be amended in ways deemed potentially viable for supporting student

learning. Adaptations may occur via change in classroom activity structure, discourse structure, group structure, or technology structure if, for example, the intended learning goals appear to be compromised. The initial HLT for this study is presented in Table 12. Consistent with the work of Simon (1995), included are the overarching learning goals, the learning activities, and the potential thinking and learning in which the in-service teachers might engage.

Rationale for the Hypothetical Learning Trajectory

According to Simon and Tzur (2004), the generation of an HLT is dependent upon the current knowledge of the students involved and the teacher's goal for student learning. A major learning goal in this study was to improve teachers' capacity to *compare distributions* using a statistical perspective. Teachers' knowledge of *comparing distributions* at the beginning of the study was approximated using the written pre-assessment of all teachers and interviews with a subset of teachers. The results of the pre-assessment and initial interviews will be discussed more fully in Chapter V. A summary is presented here.

Summary of Pre-Assessment Responses

The background and performance on the written pre-assessment suggests that high school mathematics teachers' knowledge of comparing distributions was weak. The number of statistics courses taken in college did not discriminate with respect to performance on the content assessment. That is, teachers with no formal statistics coursework did not perform significantly differently on average than their peers having taken one or more statistics courses. This sample of teachers demonstrated a collective

Table 12

Explication and Chronology of the Initial Hypothetical Learning Trajectory

Overarching Learning Goals: 1) Teachers will improve their understanding of <i>comparing distributions</i> by learning to attend to shape, center and spread in distributions as related to context and revealed in multiple representations. 2) Teachers will begin to understand the power of simulation and resampling approaches to informal inference. 3) Teachers will learn to use the language of statistics to formulate statistical questions, collect data, analyze data, and interpret results. 4) Teachers will develop facility with statistical tools including graphing calculators, <i>Fathom2</i> software, and <i>CPMP-Tools</i> software to conduct statistical explorations and analyses. <i>Potential contributions</i> to statistical big ideas are represented in column four as D=distribution; V=variability; S=sampling distributions, I=informal inference.			
Day	Learning Activity	Potential teacher (as student) thinking and learning	Big Ideas
1	Activity 1.1 Welcome and Review of Professional Recommendations and Standards for Statistics	Set the stage and create awareness of the elevation of statistics in the high school curriculum; challenge current practice and induce disequilibrium	
	Activity 1.2 Orbital Express Part I	Teachers will investigate the design of an experiment, generate numerical and graphical representations of data, compare their representations and conclusions to others, establish preliminary norms for making statistical arguments, and begin to develop some familiarity with the statistical process. The basis for whether differences between two distributions are significant is established.	D,V
	Activity 1.3 Matching Plots to Variables	Teachers will develop and communicate a shared understanding of how to construct and interpret boxplots, stem-and-leaf displays, and histograms, and consider the impact of context on the shapes of these distributions across representations.	D,V
	Activity 1.4 Standard Deviation and Its Interpretation	Teachers will begin to understand multiple measures of variability, e.g., range, IQR, SAD, MAD, <i>SD</i> , variance, and their relative strengths and weaknesses as measures. They will begin to have a sense about what is meant by bias and the reason for dividing by $n-1$ for a sample standard deviation. The relationship of <i>SD</i> to the normal distribution is introduced. The need for technology (<i>CPMP-Tools</i> , <i>Fathom2</i> , & graphing calc.) is established. Teachers will be able to compare and match statistics to distributions represented as histograms and boxplots.	D,V

Table 12—Continued

Day	Learning Activity	Potential teacher (as student) thinking and learning	Big Ideas
	Activity 1.5 Orbital Express Parts II & III	Teachers will begin to explore the essence of the randomization test physically and then with the automated randomization distribution feature of the <i>CPMP-Tools</i> software. They will use their data from Activity 1.2. It is expected that their arguments will become somewhat more sophisticated now that language of shape, centers, and variability has been explored. They will begin to think in terms of p -values and significant difference, but not formally.	D,V,S,I
2	Activity 2.1 <i>CPMP-Tools</i> —Using the Randomization Test and Estimating Mean and Standard Deviation Features	Teachers develop familiarity and facility with <i>CPMP-Tools</i> to use the built-in balancing histogram, estimate SD , and randomization distribution features. Through structured “play,” teachers will further their understanding of shape, centers, and spread. The randomization test will come more into focus.	D,V,S,I
	Activity 2.2 <i>Fathom2</i> Tours 1, 2, & 6	Teachers will now transfer to a “landscape-type” software environment and develop facility with <i>Fathom2</i> . Tours 1, 2, and 6 provide enough direction to allow teachers to navigate in the environment and to accomplish graphical and computational tasks.	D,V,S,I
	Activity 2.3 Random Walk and Orbital Express Part IV	Teachers have an opportunity to create a new case table, establish new cases with the use of a formula, re-randomize the data collection, practice graphing, collecting measures, and interpreting experimental data in a probabilistic setting. They will apply this to generating a mechanism for the randomization distribution (making transparent the multiple tiers of a sampling distribution). Comparisons between mechanisms of physical randomization, automated randomization, and constructed randomization will be made. Understanding of p -values and significant difference continue to evolve.	D,V,S,I
	Activity 2.4 Random Rectangles & Stringing Students Along	Teachers’ perceptions will be challenged via the “gorilla video” and the idea of bias reinforced. Further experience with sampling issues will support the need for random sampling, the benefit of larger samples, and the creation of sampling distributions of the mean of samples and informal introduction to standard error.	D,V,S
	Activity 2.5 <i>CPMP</i> Course 3 Unit 1: Reasoning and Proof	Teachers will be introduced to concepts of design of experiments that will allow one to confidently determine whether attribution may be determined. The second investigation will reinforce the randomization test procedure that has begun to be developed.	D,V,S,I
3	Activity 3.1 Sharing <i>Fathom2</i> Experiences and Trouble-shooting Technology Issues	Teachers will share technology issues and questions they have been experiencing and have opportunities to ask questions of instructor and others. Simulation demonstrations may further impact ideas of sampling distributions.	D,V,S,I

Table 12—Continued

Day	Learning Activity	Potential teacher (as student) thinking and learning	Big Ideas
	Activity 3.2 Seattle Real Estate Task	Teachers will continue to develop facility with <i>Fathom2</i> and the randomization test through comparison of two skewed distributions. Using means and medians as measures of center and testing for significant differences will provide a contrast and further reinforcement for the need for measures that make sense in the context of the problem. Predictions are questioned and the issue of the magnitude of a difference arises. What does significantly different mean?	D,V,S,I
	Activity 3.3 Other Randomization Test Applications	Teachers will use the randomization test while investigating different contexts and measures, e.g., matched pairs, ratios, correlation coefficients, to further understand its usefulness and broad applicability. The flexibility afforded in <i>Fathom2</i> is useful here.	D,V,S,I
	Activity 3.4 Is There a Relationship Between the Sample Size and the Sampling Distribution's Standard Deviation?	Teachers will use <i>Fathom2</i> to generate or the Internet to find and import a population in some context from which to sample. Patterns of center, variability, and shape of the sampling distribution of the mean will become evident as sample size increases. Teachers will model the relationship between sample size and standard error and conjecture as to an appropriate model. The Central Limit Theorem is “discovered” by the teachers and articulated in their words.	D,V,S
	Activity 3.5 CPMP Course 3 Unit 4: Samples and Variation	Teachers will further explore characteristics of the normal distribution and its density function. Connections to previous investigations and the “normal-looking” distributions that have been arising will be reinforced. The use of sliders in <i>Fathom2</i> to model functions will be demonstrated with the normal curve and with potential to connect to other function families—connection to function parameters.	D,V
4	Activity 4.1 Share and Summarize—Comparing Distributions: What Do We Understand?	Through conversation, a shared understanding of the ideas on the table for the week emerges explicitly. Connections to the assigned readings will be encouraged.	D,V,S,I
	Activity 4.2 Physicians' Reactions to Patient Size	This activity will encourage connections between experimental design issues and those of significant differences. It is likely that teachers' predictions of whether the difference in means or medians would be considered significant will provoke a bit of a surprise reaction. This task is intended to challenge teachers' ideas of significant differences using representations of boxplots, dotplots, histograms, and numerical summaries.	D,V,S,I

Table 12—Continued

Day	Learning Activity	Potential teacher (as student) thinking and learning	Big Ideas
	Activity 4.3 Assessing Students' Responses to Authentic Data Analysis Problems (this task was included to connect the work of learning statistics to the work of teaching statistics—looking at student work)	This activity should promote teachers' reflection on the work of students in a statistical data analysis situation. By comparing student solutions from a variety of data analysis problems, teachers may connect some of their previous struggles to those of students and become more prepared to support student understanding of statistical ideas and connect to the practice of teaching.	D,V
	Activity 4.4 Matching Samples to Density Curves	This activity will reinforce the idea that larger samples provide more information about the population. Shapes and characteristics of normal, uniform, skewed, and bimodal density curves and samples from them will be explored. Teachers will further reflect on the differences between samples and sampling distributions.	D,V,S,I
	Activity 4.5 Making Connections—From Sample to Population	Teachers will be encouraged to articulate what they believe they can tell from a single sample and connect to all of the investigations of the four days. Conversation will likely foreshadow confidence intervals.	D,V,S,I
	Activity 4.6 Closure and Closing Comments	Teachers will take a binder walk back through the activities and resources provided. Collective reflections on the accomplishments and hard work of the week will be shared. Teachers' visions of the promise and challenge of statistics in the curriculum will be discussed. Encouragement for the inclusion of statistics into the high school curriculum will be provided.	

inability to answer many assessment items adequately. Many questions on the pre-assessment were left blank or with notes such as, “I don’t remember,” “I have no idea,” or “I don’t teach this stuff.” The assessment questions on which the group performed relatively high were those associated with comparing histograms, boxplots, and dotplots, suggesting some familiarity with interpreting these representations. The items with which teachers exhibited the most difficulty included those requiring some coordination of concepts related to the Central Limit Theorem and those asking about characteristics of distributions.

Summary of Initial Interview Responses

Of the nine teachers interviewed prior to the professional development intervention, only two claimed to have positive experiences in statistics courses in college. Challenges specifically mentioned by the teachers in their statistics courses included difficulty understanding the professor due to language issues, class notes that were hand-written and not well organized, lacking pre-requisite calculus knowledge, and fast pace of the class. There was no use of technology in the courses these teachers experienced. When asked about what s/he remembered from college statistics, the following teacher's response is generally representative of seven of nine teachers (T54):

- SM: *Okay, and if you think back on that experience, are there things that you can remember from that . . . are there?*
- Lorraine: *I remember it was really fast and furious, you know, it was just thrown as a semester course that was just thrown at you.*
- SM: *Okay.*
- Lorraine: *And it was pretty easy math, but there was just so much to remember, you know, I think probability and statistics is pretty easy, like the actual crunching of numbers . . .*
- SM: *Okay.*
- Lorraine: *. . . is pretty easy, but there's . . . it's just . . . it was just so much, you know, with all the different tests, and when you use this, and how you do this, and all this . . .*
- SM: *Okay. So as far as . . . as far as thinking back on it, are there . . . are there particular things that you feel like, "You know what? I remember that really well."*
- Lorraine: *No, mean, median, mode, you know, simple. I remember like T-squared, D-squared, you know, stuff I don't . . . Chi-squared. Don't ask me when I would use them, you know, I know I could pick up a book and . . .*

This teacher had a "fast and furious" statistics course experience with potentially little useful residue. Seven other teachers' stories were remarkably similar. Only one teacher talked about activities in the course such as sampling activities or modeling activities; the remaining teachers discussed taking notes, doing homework, taking quizzes and tests.

When asked about statistics that every high school student should understand, each teacher indicated that all students should know mean, median, mode, and how to read graphs before leaving high school. With the exception of the one statistics teacher in the group, the remaining teachers suggested that students in their schools would typically not receive statistics instruction beyond that which might briefly appear in an algebra course.

When asked to compare distributions as histograms, boxplots, and dotplots, teachers generally reasoned from the range of data and some type of center, which may have included clumps, peaks, clusters, means, or medians. Shape was a characteristic that was mentioned by teachers through language of skewed, normal, or curvy shape, although many teachers did not convey confidence in their responses. The concept of statistically significant differences was not well-developed for this group as most mentioned issues of the magnitude of differences or the context of the problem in conjunction with the magnitude of differences as their rationale for determining when differences would be considered significant. Boxplots were mentioned as the least familiar representation for these teachers and several acknowledged teaching themselves about boxplots from their textbook. Though they could construct and talk about the quartiles of a boxplot, teachers struggled with interpreting boxplot information in particularly useful ways.

When referring to variability, these teachers communicated a variety of ways of understanding, but their responses did not suggest a well-coordinated view from a statistical perspective. One teacher (T16) with a fairly strong explanation spoke of variability as “diversity.” Less variability would mean greater conformity and he used language of “how far it is away from the central tendency, from the middle of the road,”

as he spoke of standard deviation, but could not recall how to calculate it. Teachers referred to range, IQR, rate of change, normal distributions, symmetry, and bumpiness when comparing two distributions with respect to variability. The issue of differing sample sizes across distributions lurked in the background for a few teachers, but was not a strong component of teachers' arguments. The interviews further suggested that teachers' ideas of the statistical process failed to attend to formulating a question from which data collection, analysis and interpretation might follow. When asked to respond to a task requiring the interpretation of results from a simulation to model a real world phenomenon, teachers in this group struggled to convey an understanding of how the simulation results may be useful. Overall, the interviews provided a sense that these teachers had some exposure to statistical ideas and language, seemed to have an informal sense of how to think about shape, center, and spread, but had very little confidence in their reasoning.

Designing the Hypothetical Learning Trajectory

Collectively, the preliminary results supported the need for learning activities with the potential to build upon some basic competencies associated with reading and interpreting graphs in light of variability, expanding the shared understanding of statistical vocabulary and use of tools, and providing opportunities for teachers to wrestle with conceptions of sampling distributions and their value to the statistical process. The HLT was constructed in light of this information.

Table 13 depicts the spiraling design of the big ideas within the learning activities from the HLT. For a more detailed illustration of the particular connections between

learning activities and big ideas of comparing distributions, see Appendix H. Each of the subcomponents of the big ideas of distribution, variability, and sampling distributions had the potential to be developed through investigations across multiple activities. There were 21 learning activities planned (19 actually implemented) for the four-day professional development program and each subtopic of each of the big ideas of distribution, variability, and sampling distribution was scheduled to be addressed in at least one learning activity per day. Each learning activity was comprised of between one and four related tasks.

Table 13

Number and Nature of Learning Activities with Potential to Support the Development of Big Ideas

Comparing Distributions: Big Ideas	Number of Activities Potentially Supporting the Development of Big Ideas (19 total activities)				
	Day 1	Day 2	Day 3	Day 4	Total
Distribution					
Characteristics of distributions	4	4	4	5	17
Comparing dot plots	4	5	4	4	17
Comparing histograms	4	5	3	5	17
Comparing box plots	4	5	2	3	14
Variability					
Context	4	3	3	3	13
Graphical representations	4	4	4	4	12
Sample size	3	1	1	2	7
Sampling Distributions					
Central Limit Theorem	1	1	1	2	5
Simulation	3	3	3	2	11
Design of experiments	2	2	2	4	10

One consequence of this design structure was that in the event of time constraints or some other emergent impeding condition during the professional development program making the completion of all 21 activities impossible, the significant spiraling and overlapping of big ideas would potentially still enable teachers to construct an improved understanding of *comparing distributions*. The entire experience was designed

with sensitivity to the learning principles of engaging and building on the learners' prior understandings, attention to the development of, and connections between, conceptual and procedural understanding, and opportunities for self-reflection on the learning experiences (National Research Council, 2005). Additionally, as the study was situated in the context of statistical professional development for high school mathematics teachers, issues of professional development and statistical content were important considerations in the design and implementation of the intervention.

Professional Development Considerations

The high school mathematics teachers who participated in this study all teach in rural school districts, use instructional materials that may be considered “non-progressive” in nature, use technology to a limited degree with students, and have historically attended only minimally to statistics in the high school curriculum. These contextual factors are important considerations in the design of professional development as one attempts to engineer learning activities with potential to engage teachers and build on their current statistical understandings. The design of the professional development required sensitivity to a host of competing factors:

1. Perceived content need of the teachers by the teachers and by the researcher;
2. Perceived pedagogical content knowledge need of the teachers by the teachers and by the researcher;
3. Time—four six-hour days (approximately 20 instructional hours) to conduct the session;

4. Recognition that the curriculum materials for the professional development program may or may not be seen as relevant to the teachers.

The umbrella-project supporting the current study was a two-year project designed to support teachers' mathematical content knowledge for teaching. It was not the case that all districts in the project were using common student instructional materials or would ever be using common instructional materials. Also, the instructional materials used in each of the districts within the project at the time of this study, would be considered conventional materials. Textbooks generally supported the mathematical practice of teacher-led instruction, including a number of worked-out examples, followed by student practice. Lesson observations conducted as part of the larger project strongly supported the view that classrooms of these teachers were mainly teacher-centered, procedurally-dominated instructional venues. The umbrella-project's professional development sessions were designed to provide opportunities for teachers to experience mathematics from a more problem-oriented perspective and to model instructional practices that might be associated with such an approach with high school students. The research presented in the present study assumes a similar posture.

Statistical Content Considerations

Teachers' evolving understanding of comparing distributions was a focus of this study. As represented by Figure 2 at the end of Chapter II, *comparing distributions* is broadly conceived to encompass a triadic, multidirectional relationship between the statistical ideas of distribution, variability, and sampling distributions. The following

sections will discuss the importance of attending to variability in the HLT as well as issues of task selection and sequencing.

Variability as fundamental to understanding comparing distributions. Variability, as a watershed statistical concept, is fundamental to understanding comparing distributions (Gould, 2004; Moore, 1990). Variability plays a role both within distributions and between distributions. It plays an extremely important role in the study of sampling distributions (Pfannkuch, 2006). Because of the inherent complexities associated with understanding variability and our limited collective knowledge of how that understanding develops, attending to variability was an important consideration in the construction of the learning sequence for teachers. Drawing on theories of constructivist learning (Noddings et al., 1990; M. A. Simon, 1995; vonGlaserfeld, 1990, 1995), Garfield and Ben-Zvi (2005) argue that

progress in students' construction of meanings is not linear but rather complex and is best captured by the image of spiral progression. Therefore, ideas related to variability must be constantly revisited along the statistics curriculum from different points of view, context and levels of abstraction, to create a complex web of interconnections among them. (p. 95)

With *comparing distributions* as the content focus, statistical covariation was not a focus of the present study. Thus, the variability reasoning frameworks presented in Chapter II in the section on “Students’ and Teachers’ Understanding of Variability” were adapted to this study. In particular, opportunities to develop knowledge of variability including “identifying patterns of variability in fitting models” (Garfield & Ben-Zvi, 2005), or “as covariation or association” (Shaughnessy, 2007) were not built into the HLT. However, the non-covariational aspects of the frameworks were explicitly attended to in the HLT.

Additional direction for the development of the HLT was based upon the work of Pfannkuch et al. (2004). These researchers found that 15-year old students in New Zealand had difficulty drawing conclusions when comparing data plots. They hypothesized that the curriculum in its present form did not sufficiently support the transition between informal and formal inferential thinking. By analyzing student assessment task responses and classroom learning opportunities, they devised a framework which they suggested may help improve student success with comparing data plots. In particular, the recommended framework was designed to support teachers' work with students on inferential reasoning and includes "(i) knowing why they should compare centres, (ii) describing and interpreting variability within and between sample distributions, (iii) developing their sampling reasoning, and (iv) how to draw an acceptable conclusion based on informal inference" (p. 5). Teachers in the current study exhibited similar difficulties at the time of the pre-assessment as the 11th-grade students in Pfannkuch and colleague's study. Thus, their framework was used to further inform the development of the HLT.

The aforementioned considerations in conjunction with the literature reviewed in Chapter II provided the basis for the development of the initial HLT. Still, decisions for the selection or creation of tasks and the sequencing of the learning activities involved myriad choices. Learning activities as described in Table 12 were frequently composed of multiple tasks. The following design principles guided the task selection or creation and sequencing.

Task selection. Tasks were selected or created base upon their potential to:

1. engage teachers and create cognitive dissonance for teachers about existing statistical conceptions. Tasks requiring teachers to make sense of representations were privileged over those exclusively requiring the construction of representations.
2. support the development of, and connections between, big ideas of statistics. It was conjectured that through investigating the use and workings of the randomization test, manually and with the use of technology, teachers' ability to compare distributions would improve.
3. support the use of technology. Using large data sets and resampling methods made use of technology essential to investigations. One hypothesis was that through investigating the tasks with technology, teachers might develop facility with the tool as well as begin to understand important underlying statistical principles.
4. support patterns of discourse and the development of sociomathematical norms that would potentially facilitate learning of statistics and model what might be possible with students.

Many of the tasks and investigations identified as promising for use in the HLT were selected from existing sources and organized in conjunction with the design principles described above. Additionally, several activities were modified or created by the researcher with the express purpose of challenging teachers' ideas of statistics and connecting other mathematical areas with which teachers may feel more competent.

Task sequencing. Systematically-structured statistical tasks provided the basis for the ongoing collegial problem-solving and dialogue in the professional development

program. The sequence of tasks emerged from the attention to the design principles described above. Specifically, the decision was made to begin the session, after a brief introduction to the topic focus and associated professional recommendations for statistics in the secondary curriculum, with a task that might be problematic for teachers, yet embody the essence of a statistical process (Wild & Pfannkuch, 1999). It was hypothesized that the experience would create “a need to know” for the teachers and subsequently a series of additional investigations might support or scaffold teachers’ understandings and promote conceptual and procedural understanding of comparing distributions and other closely-associated big ideas. Hypothetically, teachers would reconcile the problematic nature of the original task and resolve the initial disequilibrium through the construction of powerful statistical understanding and connections.

This approach is consistent with other researchers and in particular, potentially representative of an attempt to engineer “productive disciplinary engagement” (Engle & Conant, 2002). Engle and Conant state that “productive disciplinary engagement can be fostered by designing learning environments that support (a) problematizing subject matter, (b) giving students authority to address such problems, (c) holding students accountable to others and to shared disciplinary norms, and (d) providing students with relevant resources” (p. 399). Although their work was in the context of elementary school students, it was hypothesized that productive disciplinary engagement may well apply to professional development with teachers.

Appendix G contains the original instructor guide pages for each of the four days of the professional development program. Because the researcher and the instructor are the same in this study, the guide pages were written with somewhat less detail than one

might anticipate for guide pages to be used by another person. Since many of the activities and investigations were selected and adapted from existing sources, further instructor support may be found directly within those sources.

Relationship Between the HLT and Assessment

Written pre- and post-assessments of statistical content knowledge (see Appendices A and B) were designed and administered to teachers to attempt to partially determine change in teachers' understanding of *comparing distributions*. Each item on the assessment instrument was designed to provide opportunity for teachers to demonstrate their understanding of distributions, variability, and sampling distributions. Figure 5 provides a visual representation of the associations between the big statistical ideas of distribution, variability, and sampling distribution in this study and the written content assessment items. Appendix H illustrates the association between the planned learning activities from the HLT and the assessment items in this study.

The illustrations in this chapter (see Tables 12, 13, and Figure 5) provide a sense of the overall design involved in this research. It should be apparent that neither the HLT nor the written pre- and post-assessment suggests a linear approach to the development of these concepts. Instead, teaching and assessing the understanding of *comparing distributions* incorporated a complex array of tasks and assessment items with substantial overlap with respect to the related concepts of distribution, variability, and sampling variability. The arrows in the tables signify potential relationships between items; big ideas are organized roughly according to increasing level of perceived potential statistical difficulty or complexity. The increasing complexity may be thought of as increasing

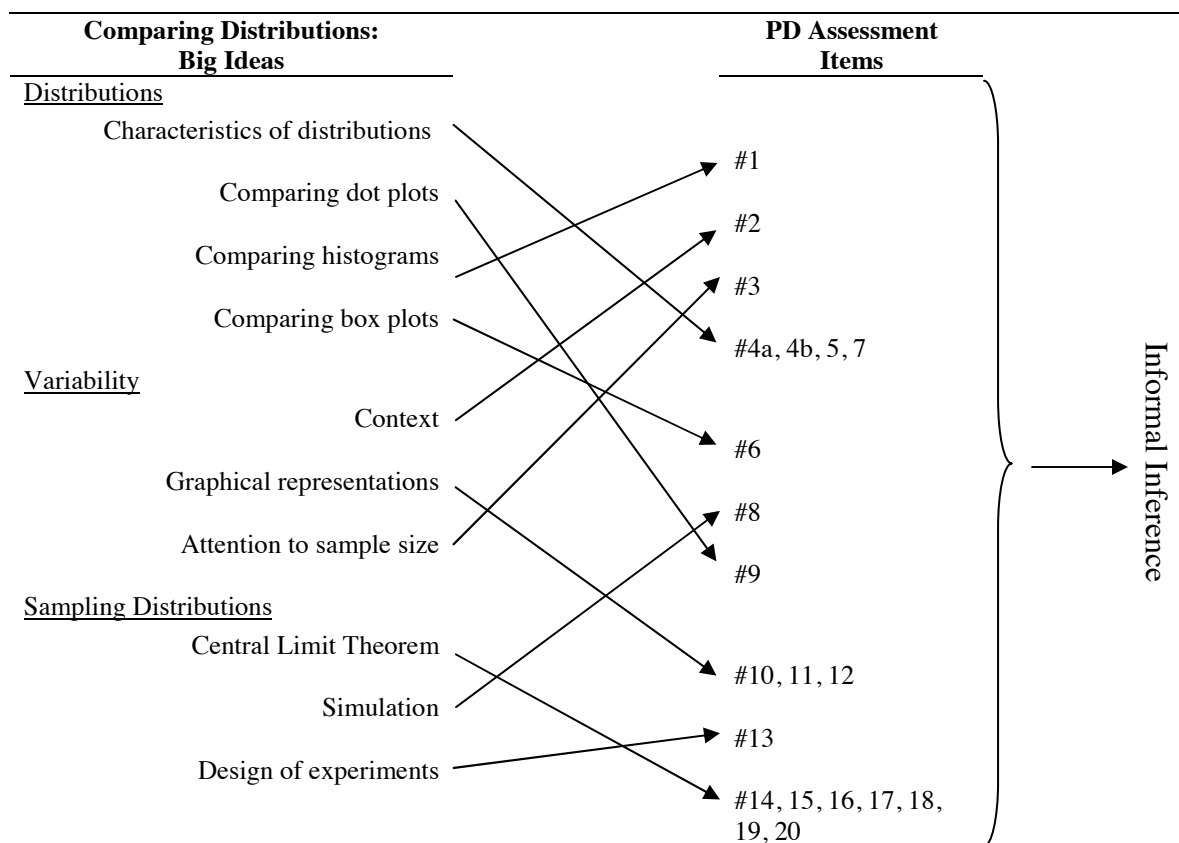


Figure 5. Mapping content to assessment.

cognitive demand (Stein, Smith, Henningsen, & Silver, 2000). As the learning experiences were co-constructed across three professional development venues, modifications to the HLT were made. The transformation from the designed or envisioned HLT to the enacted learning trajectory is described next.

The Emergent Learning Trajectory

The learning trajectory was investigated across three distinct professional development settings. Though each of the settings was similar, the constellation of teachers and facilities across groups were quite variable. Group sizes ranged from 31 to 9 to 16 teachers, creating differences in conversations and dynamics. The availability of

computer technology ranged from laptops on tables at all times, to needing to move to a computer lab for select activities, to some teachers with laptops and some in a computer lab. These differences posed logistical challenges for the researcher and also provided realistic classroom environments in which to test an HLT. As will be discussed more fully in Chapter VI, the variation in conditions across groups and the resulting similarities in conversation and assessment results support the robustness of the HLT.

The initial HLT was presented at the beginning of this chapter. As the learning trajectory was enacted during the first professional development program, a number of modifications were required. First, because the group was fairly large ($n = 31$), dialogue and whole group conversations lasted longer than anticipated. Because of the richness of the conversations, it was decided that the potential benefit of worthwhile collegial dialogue outweighed the cost of not completing all activities. Additionally, due to some technological difficulties with the use of software on Day 1, some of the technology-oriented investigations were briefly postponed. Ultimately, two investigations were omitted, Activities 3.3 and 4.3. Handouts were provided in the participant binder for the two activities, but no class time was devoted to either. Because of the size of this group, to complete all of these tasks in four days was challenging and required the researcher and the teachers to remain on task and focused throughout the duration of the professional development experience. With the exception of scheduled breaks and lunch during the day, there was almost no down-time during the four-day session. Every teacher appeared to engage in the activities and work diligently as evidenced by both the classroom observations during the sessions and the videotape analysis of the activity.

Tables 14, 15, and 16 display the sequence of learning activities as enacted during the first, second, and third professional development programs, respectively. Activities whose sequence deviated from the initial HLT are notated with underlining and can also be identified by the activity number out of sequence; activities that were eliminated due to time constraints are lined out, but the content is preserved for comparison. Brief notes or explanations for some activities are found in bold capital letters.

Though not in identical fashion, teachers at all three professional development sites had the opportunity to engage in a similar learning experience. A comparison across sites shows the same set of 19 activities was experienced by all groups. The opportunity to learn statistics was not significantly different across sites by virtue of the set of activities with which teachers could engage. However, as with any set of learning environments, a number of variable forces may have impacted the enacted learning trajectory. As can be seen by comparing Tables 14, 15, and 16, there are some minor modifications to the order of activities across sites. The depth or direction of conversations that were pursued as teachers' constructed statistical justifications differed across sites due to the differences in participants and group dynamics, but the opportunity to engage with and make sense of statistical content and technology was remarkably similar across groups.

To illustrate the nature of the learning environment and discourse structure, the following section provides selected transcripts from the videotape of the professional development programs. The selections were made on the basis of their potential to serve as paradigm cases of the nature of the co-constructed learning environment as well as for their possible benefit from an analytical viewpoint. During the retrospective analysis,

Table 14

Learning Activities as Sequenced and Enacted in PD Session 1

Day	Learning Activity	*
1	Activity 1.1 Welcome and Review of Professional Recommendations and Standards for Statistics	
	Activity 1.2 Orbital Express Part I	D,V
	Activity 1.3 Matching Plots to Variables	D,V
	Activity 1.4 Standard Deviation and Its Interpretation THIS IS WHERE DAY 1 ENDED	D,V
2	Activity 1.5 Orbital Express Parts II & III	D,V,S,I
	Activity 2.1 <i>CPMP-Tools</i> —Using the Randomization Test and Estimating Mean and Standard Deviation Features	D,V,S,I
	Activity 2.2 <i>Fathom2</i> Tours 1, 2, & 6 THIS ACTIVITY BEGINS AT 11:30 ON DAY 2	D,V,S,I
	Activity 2.4 Random Rectangles & Stringing Students Along	D,V,S
	Activity 2.5 CPMP Course 3 Unit 1: Reasoning and Proof	D,V,S,I
3	Activity 3.1 Sharing <i>Fathom2</i> Experiences and Trouble-shooting Technology Issues	D,V,S,I
	Activity 2.3 Random Walk and Orbital Express Part IV MOVED TO DAY 3	D,V,S,I
	Activity 3.2 Seattle Real Estate Task	D,V,S,I
	~~Activity 3.3 Other Randomization Test Applications~~ SKIPPED THIS ONE FOR TIME	~~D,V,S,I~~
	Activity 3.4 Is There a Relationship Between the Sample Size and the Sampling Distribution's Standard Deviation? TRIED TO PUSH FOR THIS AND IT WAS WAY TOO MUCH—CONFUSION AND FRUSTRATION WERE EVIDENT	D,V,S
	Activity 3.5 CPMP Course 3 Unit 4: Samples and Variation	D,V
4	Activity 4.1 Share and Summarize—Comparing Distributions: What Do We Understand?	D,V,S,I
	Activity 3.4 Is There a Relationship Between the Sample Size and the Sampling Distribution's Standard Deviation? REVISITED THIS ACTIVITY SUCCESSFULLY	D,V,S
	Activity 4.2 Physicians' Reactions to Patient Size	D,V,S,I
	~~Activity 4.3 Assessing Students' Responses to Authentic Data Analysis Problems~~ (this task was included to connect the work of learning statistics to the work of teaching statistics—looking at student work) SKIPPED THIS ONE FOR TIME	~~D,V~~
	Activity 4.4 Matching Samples to Density Curves	D,V,S,I
	Activity 4.5 Making Connections—From Sample to Population	D,V,S,I
	Activity 4.6 Closure and Closing Comments	

Table 15

Learning Activities as Sequenced and Enacted in PD Session 2

Day	Learning Activity	*
1	Activity 1.1 Welcome and Review of Professional Recommendations and Standards for Statistics	
	Activity 1.2 Orbital Express Part I	D,V
	Activity 1.3 Matching Plots to Variables	D,V
	Activity 1.4 Standard Deviation and Its Interpretation THIS IS WHERE DAY 1 ENDED (25 minutes during Day 2 to debrief)	D,V
2	Activity 1.5 Orbital Express Parts II & III	D,V,S,I
	Activity 2.1 <i>CPMP-Tools</i> —Using the Randomization Test and Estimating Mean and Standard Deviation Features	D,V,S,I
	Activity 2.2 <i>Fathom2</i> Tours 1, 2, & 6 THIS ACTIVITY BEGAN AT 11:30 ON DAY 2	D,V,S,I
	Activity 2.4 Random Rectangles & Stringing Students Along	D,V,S
	Activity 2.5 CPMP Course 3 Unit 1: Reasoning and Proof	D,V,S,I
3	Activity 2.5 is Debriefed	
	Activity 3.1 Sharing <i>Fathom2</i> Experiences and Trouble-shooting Technology Issues	D,V,S,I
	Activity 2.3 Random Walk and Orbital Express Part IV MOVED TO DAY 3	D,V,S,I
	Activity 3.2 Seattle Real Estate Task	D,V,S,I
	~~Activity 3.3 Other Randomization Test Applications~~ SKIPPED THIS ONE FOR TIME	~~D,V,S,I~~
	Activity 3.4 Is There a Relationship Between the Sample Size and the Sampling Distribution's Standard Deviation? INTRODUCED THIS USING A RANDNORM(54,2) POPULATION AND ACTIVELY ENGAGED TEACHERS IN THE CONSTRUCTION OF THE SAMPLING DISTRIBUTION MECHANISM. ALL TEACHERS GENERATED POPULATION AND MECHANISM AND COMPLETED TABLE; MODEL WAS POSTPONED UNTIL DAY 4.	D,V,S
	Activity 3.5 CPMP Course 3 Unit 4: Samples and Variation INTRODUCED AND ASSIGNED (Inv. 1 #1-6, Inv. 2 #1)	D,V
4	Activity 3.5 CPMP Course 3 Unit 4: Samples and Variation SOLUTIONS ARE COMPARED AND SHARED UNDERSTANDING PURSUED	
	Activity 4.1 Share and Summarize—Comparing Distributions: What Do We Understand?	D,V,S,I
	Activity 3.4 Is There a Relationship Between the Sample Size and the Sampling Distribution's Standard Deviation? REVISITED THIS ACTIVITY SUCCESSFULLY	D,V,S
	Activity 4.5 Making Connections—From Sample to Population	D,V,S,I
	Activity 4.2 Physicians' Reactions to Patient Size	D,V,S,I
	~~Activity 4.3 Assessing Students' Responses to Authentic Data Analysis Problems (this task was included to connect the work of learning statistics to the work of teaching statistics—looking at student work)~~ SKIPPED THIS ONE FOR TIME	~~D,V~~
	Activity 4.4 Matching Samples to Density Curves	D,V,S,I
	Activity 4.6 Closure and Closing Comments	

Table 16

Learning Activities as Sequenced and Enacted in PD Session 3

Day	Learning Activity	*
1	Activity 1.1 Welcome and Review of Professional Recommendations and Standards for Statistics	
	Activity 1.2 Orbital Express Part I CLOSED USING H-14, CPMP STUDENT QUIZ SCORES ACTIVITY—THIS IS DIFFERENT THAN SESSION 1	D,V
	Activity 1.3 Matching Plots to Variables DID NOT USE <i>FATHOM2</i> TO DEMO DISTRIBUTIONS AND IMPACT ON MEDIAN AND MEAN BECAUSE OF TECHNOLOGY ISSUE	D,V
	Activity 1.4 Standard Deviation and Its Interpretation USED MULTIPLE MEASURES OF VARIABILITY TO LOOK FOR RELATIVE RELATIONSHIPS. <i>FATHOM2</i> GETS INTRODUCED HERE (AND MEAN/MEDIAN ISSUE IS RESOLVED) ALSO USED <i>FATHOM2</i> TO ILLUSTRATE HANDSPANS $\pm$ SD(HANDSPANS). INTRODUCED <i>CPMP-TOOLS</i> BALANCING HISTOGRAM & SD ESTIMATOR	D,V
2	<u>Activity 1.4 continues.</u> DEBRIEFED H-31 AND DISCUSS H-32 (Properties of estimators) FINISHED WITH MATCHING STATISTICS TO PLOTS ACTIVITY H-33-36 REFLECTION #1	
	<u>Activity 1.5</u> Orbital Express Parts II & III DID MUCH OF THIS AS A GROUP	D,V,S,I
	Activity 2.1 <i>CPMP-Tools</i> —Using the Randomization Test and Estimating Mean and Standard Deviation Features DEMOED THIS EARLIER	D,V,S,I
	Activity 2.2 <i>Fathom2</i> Tours 1, 2, & 6 THIS ACTIVITY BEGAN AT 11:45 ON DAY 2 NEEDED EVERYONE TO COMPLETE TOURS 1 & 2 BEFORE LUNCH WAS OVER; TOUR 6 IF POSSIBLE	D,V,S,I
	<u>Activity 2.3</u> Random Walk and Orbital Express Part IV SHIFTED BACK FROM DAY 3 DURING SESSION 1 TO DEMO AND MODEL FOR PARTICIPANTS TO CONSTRUCT DAY 2 REFLECTION #2	D,V,S,I
	Activity 2.4 Random Rectangles & Stringing Students Along	D,V,S
	Activity 2.5 CPMP Course 3 Unit 1: Reasoning and Proof THIS ACTIVITY WAS BEGUN AND #3 & 4 FROM INVESTIGATION 2 WERE ASSIGNED FOR HOMEWORK.	D,V,S,I
3	<u>Activity 2.5 is debriefed</u>	
	Activity 3.1 Sharing <i>Fathom2</i> Experiences and Trouble-shooting Technology Issues	D,V,S,I
	Activity 3.2 Seattle Real Estate Task	D,V,S,I
	~~Activity 3.3~~ ~~Other Randomization Test Applications~~ SKIPPED THIS ONE FOR TIME	~~D,V,S,I~~
	Activity 3.5 CPMP Course 3 Unit 4: Samples and Variation WORKED ON INVESTIGATION 1 DURING SESSION, ASSIGNED INVESTIGATION 2, #1-6 TO BE COMPLETED FOR HOMEWORK. REFLECTION #4	D,V

Table 16—Continued

Day	Learning Activity	*
4	Activity 3.5 is debriefed FATHOM2 FUNCTION PLOTTING IS DEMOED AND DISCUSSED	
	Activity 4.1 Share and Summarize—Comparing Distributions: What Do We Understand? THIS ACTIVITY WAS ESSENTIALLY ELIMINATED; DISCUSSION WAS INCORPORATED INTO REMAINING ACTIVITIES	D,V,S,I
	Activity 3.4 Is There a Relationship Between the Sample Size and the Sampling Distribution's Standard Deviation?	D,V,S
	Activity 4.4 Matching Samples to Density Curves MOVED AHEAD OF ACTIVITY 4.2 TO ASSIST TEACHERS WITH MAKING CONNECTIONS	D,V,S,I
	Activity 4.5 Making Connections—From Sample to Population NOT REALLY AN ACTIVITY, JUST A DISCUSSION REFLECTION #5	D,V,S,I
	Activity 4.2 Physicians' Reactions to Patient Size	D,V,S,I
	~~Activity 4.3 Assessing Students' Responses to Authentic Data Analysis Problems~~ ~~(this task was included to connect the work of learning statistics to the work of teaching statistics—looking at student work)~~ SKIPPED THIS ONE FOR TIME	~~D,V~~
	Activity 4.6 Closure and Closing Comments VERY ABBREVIATED	

sections of videotape were identified as representing the occurrence of potentially important transformations or noteworthy learning sequences. As themes began to emerge across time and sites, pieces of transcripts representing particular themes were identified. One difficulty that became apparent was that every one of the 19 enacted learning activities appeared to add value to the teachers' learning, thus trying to select from the 19 activities and then further refine the selection to pieces of transcripts became quite challenging. The purpose of the following section is two-fold: (1) to orient the reader to the nature of the enacted professional development beyond that of the listing of activities, and (2) to provide the basis for analysis of the enacted professional development.

Select Episodes from the Enacted Learning Trajectory

It is impossible to present 60 hours (20 hours across three sites) of transcribed professional development activity in its entirety in this medium; therefore professional development episodes are selected based on their collective potential to illuminate answers to the research questions in the study. What follows are descriptions of enacted activities in conjunction with transcripts of episodes which together provide a sense of the nature of the investigations and discourse during the sessions. The selections come from video of the professional development programs and are coded according to the site. Transcripts are coded by site, tape, and time. For example, (Site 3, Tape 1, 41:16) represents transcript from the third professional development site, tape 1, and beginning 41:16 minutes into the tape. Excerpts from each of the four days are provided to illustrate the evolution of the discourse in the group as well as the statistical content trajectory. More detail is provided for Day 1 than the other days in order to establish a general sense of the tone and direction of the session. Following the details and episodes in the forthcoming sections, a summary of major characteristics of the professional development intervention and content development will be presented.

Day 1

Setting the stage for statistical investigation and discourse. The Orbital Express Activity 1.2 was the first statistical investigation of the session. It followed the introductory activity in which teachers were exploring state and national recommendations for statistics in the high school curriculum. During Activity 1.2, teachers conducted experiments to determine which of two orbital vehicles would land

nearest a target when dropped from space. The following transcript is provided as baseline reference data for the session (Site 3, Tape 1, 41:16). SM represents the facilitator and teachers' names have been changed to pseudonyms.

- SM: *Before we start this, can you talk to me a little bit about some stuff that we should maybe attend to? [long pause—9 seconds] Well you're going to have two kinds of paper and you're going to be asked to collect seven pieces of data for each piece of paper.*
- Dillon: *You need to make sure you're dropping it from the same orbital height.*
- SM: *OK, and why would we want to do that?*
- Dillon: *So that you have a data set that is consistent across the, in terms of the physical forces, impact of the drop, consistent across all three.*
- Amy: *We want one variable, not, if we change the heights, that would be two variables.*
- SM: *OK, So you guys are kind of talking about the same thing, you're going to say let's just make one variable, did I hear that right? Or we're going to kind of control for the height, we're going to make that consistent so that it doesn't introduce a second variable.*
- Dillon: *We could do another study to see if a different location in space, when it's dropped impacts things, but it didn't sound like that was part of the study.*
- Scott: *We're trying to measure what the distance from center is, aren't we?*
- SM: *I don't know. Well we've got an issue here with the height making sure.*
- Scott: *Each one of these, that's what I was trying to . . .*
- SM: *These very scientifically-constructed vehicles here. OK, but Scott you bring up an interesting point. You said something about measuring.*
- Scott: *We're measuring the distance from center that it finishes, not necessarily where it hits . . . the shape does have something to do with it.*
- SM: *So then how are you going to then decide when it hits, assuming you have a target perhaps, let's just do it on the table so everyone can see.*
- Wilson: *What if you're not a very good aim?*
- SM: *What if you're not a very good aim? So if you mark the X or something on the floor and the target or the vehicle lands there [demonstrating], how are we going to measure that? Is there . . . should we . . .*
- Scott: *Center to center or center to closest edge.*
- SM: *OK, what do other people . . . I'm asking for advice here. What would you like to do?*
- Beatrice: *You just have to be consistent, whether it's center to center or center to edge. You've gotta make sure [inaudible].*
- SM: *Let's see if we can get everybody to agree on this. How would we like it to be?*
- Beatrice: *Center to closest edge.*
- SM: *[raising a hand to signify voting] Center to closest edge?*
- Callie: *By closest edge do you mean closest edge or . . .*

SM: *Ahh. What do you want to do there guys? Or doesn't this all matter?*

Amy: *I don't think it matters if everyone does it the same for their own.*

Beatrice: *Or do we want to be able to compare studies . . . that's why you want us to all be the same?*

SM: *I think for right now, what's going to be most important is within your group, whatever you guys do, you do it consistently. It may be the case that another group takes another approach here and that's OK, but it needs to be sort of well-defined. Whatever it was you were doing . . .*

Dillon: *You said there are three different papers but we're doing, each group is only doing two. Are you deciding which two?*

SM: *Nope, you get to pick, whichever ones you want.*

Dillon: *You want us to make a hypothesis as to what . . .*

SM: *You're going to go over there and grab whatever two papers you are interested in testing and then you're going to look at them somehow and come to an agreement amongst your group about which do you think is going to do the better job here, interpreting what does that mean. OK, so now we've got this issue of height sort of thing and it sounded like we agree that we're going to have to start from the same height . . .*

Group: *[Some question of starting height]*

SM: *The only condition that I'm going to put on you is that I want it to be higher than six feet. I don't want you to drop it from like here [motioning toward about 18 inches from the floor]. I want it to be at least a six-foot drop. After that, you decide whatever criteria that you want. OK. Are there other issues we should be on the lookout for, things that could happen?*

Stephan: *Somebody said, "What if your aim was off?" Let's say you've got one paper and say your target was right here [motioning to a location on the tabletop], and all your things landed in this nice little tight group, and your other one, your other one landed like this [motioning in a wide swath around the target], when you have this one tight group that's probably way out over here [motioning away from the target], so yeah, it's a real tight group so really your aim was off, if you could move your aim it'd be perfect. So may . . .*

Scott: *We should have a pendulum right and we should be sure we have an imaginary spot right over the top.*

Stephan: *I'm just saying, if we get all the things in a little tight group over here, it just means your aim's off, it doesn't mean your stuff's not good. This is really what you want, you just need to change your aim.*

SM: *So in the directions, it does say before you start your experiment, you may want to do a couple practice drops, because, for that exact reason Stephan. Maybe you've got something right here, but you need to adjust your starting position just a little bit, right. So that's where your quality assurance person is going to come in handy too, because you're going to want to make sure wherever you decide that starting point is, once you decide it, that, that's really it. Because once you go moving it, what happens?*

Scott: *Should we put markers down?*

SM: *Well, you're going to have to figure out some way that you're going to control for this stuff. Alright, are there any other things we should be thinking of [several clarifications about equipment and logistics]?*

SM: *[Back to tabletop example] But even this deal right here. What would happen if we measured that distance, if everybody did it in isolation? Didn't get to watch anybody else do it and we came over here and we measured. The target's right there, the vehicle's right there. Do you suppose we would all measure the same?*

Callie: *No.*

SM: *You're saying no, why not?*

Callie: *Cause it depends on the units.*

SM: *Let's even say we use the right units, maybe we all agree we're going to use centimeters. I think you all agreed you are going to go from the side of the vehicle to the . . . center of the target.*

Teacher: *So you just put a dot there [signifying the target].*

SM: *OK, so you're going to put a dot there [placing a dot on the tabletop target].*

Group: *Lots of verbal response.*

SM: *So quality assurance person, are you sensing what is part of your job here is to make sure whatever you do, you do it the same every time? That we agree on whatever it is we're doing. Any other things? [pause] You need seven of each type. Are you going to drop all seven of this one and then all seven of the next one? Is that a good idea?*

Scott: *The atmospheric conditions aren't going to change, like the wind and all.*

SM: *OK, we're OK with the atmospheric conditions, is there anything else we might need to be worried about in this instance?*

Amy: *Just how it's being dropped or how your hand is.*

SM: *So if I drop it like this [arm raised, wrist straightened] or if I drop it like this [arm raised, wrist bent], that might be different? OK. Anything else?*

Callie: *Are you saying we're using seven different sheets of paper or are . . . we're just dropping that one piece seven times?*

SM: *Dropping this one thing seven times.*

Dillon: *Are you reshaping that each time, is that what you're getting at? If we're dropping the same piece of paper, are we reshaping the vehicle to its original?*

SM: *It's your experiment, I don't know, what are you going to do?*

Teacher: *You're going to have to give it a good squeeze every time.*

SM: *You want to give it a good squeeze every time, so you're going to have somebody who's going to be the squeezer, there going to be a uniform squeezer [laughter]. Don't even get me started [more laughter]. OK, so you're going to have to deal with that.*

Callie: *I'm even thinking that the shape of it is not completely round so [using arm gestures to suggest different falling patterns] that if it falls one way, will it make it move . . .*

- Group: [Many comments]. . . *How you hold it in your hand, we should almost mark the top in it . . .*
- SM: *So you'll deal with that in some ways. Alright, should we do all seven of these and then do all seven of the others?*
- Group: *No . . . yeah.*
- SM: *Anyone take a different opinion [no response]. So you want to do seven and then do seven [wait]? OK. Well, let's see how this works out. So does everybody know what you're going to be doing . . . [A few logistical directions]*
- SM: *On page H-13, your group is going to be charged with, on poster paper, creating some visual display of your choosing, that effectively represents your data and you have to say which is the better design, based on your evidence. Identify your group roles and get to work.*

The intention during this initial investigation was to begin to establish a safe, collegial environment in which teachers would actively participate in dialogue and begin to share authority for the knowledge in the classroom environment. An examination of the discourse suggests an exchange pattern alternating from teacher to facilitator, with facilitator clarifying, revoicing, challenging, or redirecting teachers' statements. Use of humor can be seen in the interchanges. Seeds were planted during the discussion relating ideas of sampling variability, measurement variability, experimental design, and bias. None of these concepts were fully developed or resolved; the intent was to simply provoke discussion and to engage teachers' prior ideas regarding these concepts. It is evident from the transcript that at least some teachers had some prior understanding of issues of experimental design and controlling for variables in a study. Dialogue during this preliminary investigation was similar across all three sites.

Supporting statistical justification, representation, and norms of participation.

Following data generation and collection, teachers were prompted to look around the room at the posters created by the groups to try to make sense of what they were seeing (Site 3, Tape 1, 1:19:25). Each group then reported out what they did during their

experiment, their initial conjecture as to which vehicle they predicted to perform best, an explanation of their representation(s), and their conclusion. Group representations varied and included side-by-side boxplots, line plots, bargraphs comparing average distance from the target, graphs resembling targets with concentric rings, and non-graphical numerical summaries.

Boxplots were the most frequently occurring representation; however, this appeared to be somewhat of a copycat phenomenon. When groups were constructing posters, many were watching carefully what others were doing, almost as if they were not sure what would be appropriate. That boxplots were seen most often was somewhat surprising because during the pre-intervention interviews, teachers suggested they were least familiar and comfortable with boxplot representations. Most teachers said they knew how to make boxplots, but were not confident in being able to reason from them. A number of teachers mentioned that they learned about boxplots only from their school textbook and within the last 10 years.

The discourse during this poster-sharing episode allowed for pre-existing ideas and statistical language to appear, preliminary interpretations and comparisons to be made, and for groups to begin wrestling with conceptions of statistically significant differences. During the discussion, it became clear that many teachers understood “significantly different” to mean a contextual difference that was very important to consider or a large magnitude difference when the means or medians of two distributions were relatively far apart. Comparing distributions at this point involved some comparison of location via comparison of mean, median, or range; some discussion of range and IQR; and some discussion of outliers. Shape of the distributions was not mentioned; however,

the datasets contained only seven values. With the multiple experimental results posted on the walls for future reference, this discussion ended with a sense of what was to come and very tentative conclusions. This was consistent across all three sites.

The next phase of Activity 1.2 came from the draft Core-Plus Mathematics Project, Course 1, *Patterns in Data* unit (Site 3, Tape 1, 1:43:40). The activity was selected to support teachers' boxplot reasoning using a context of student quiz scores. It was conjectured that the context would be useful to encourage teachers to make data-based and belief-based arguments about grading. Twenty quiz scores for each of five students were provided in the task. Corresponding boxplots were shown for four of five students. Teachers were asked to construct a boxplot for the fifth student's (Susan) grades. Some teachers reached immediately for graphing calculators and were advised not to use the technology at that time because of the formative assessment nature of the task. Below is an illustration of one group of teachers lacking the knowledge of how to construct a boxplot and yet being willing to be vulnerable by making that lack of knowledge known. The following transcript is taken as an indication of some teachers beginning to feel intellectually safe in the environment (Site 3, Tape 1, 1:45:00).

- Megan: *Can I ask you a question?*
SM: *Sure.*
Megan: *We were discussing about our boxplot. I said that this would be a funny boxplot with all eights inside the box and she said [referring to Beatrice], well then the ends go at $7\frac{1}{2}$ and $8\frac{1}{2}$. [the scores were 6, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9]*
Beatrice: *You just average these . . .*
Megan: *Well that's my question.*
Beatrice: *Is that how you do it?*
SM: *Oh . . .*
Megan: *I would say that the box should just be from 8 to 8, but I like her suggestion.*
SM: *OK, that's good, but you're not sure?*

- Megan: *She's probably right, I'm just not as familiar with these as she is*
[referring to Beatrice].
Amy: *I didn't know how to find the quartiles without the calculator.*

This type of dialogue began to occur regularly during the professional development program and was taken to suggest a sense that teachers felt safe to declare lack of understanding to both peers and facilitator without fear of negative judgment.

During the whole group debriefing of the task, the group questioned and clarified the construction of boxplots and challenged the notion of outliers. The term “outlier” had been used previously by several groups and so this term was clarified using Tukey’s outlier definition of values smaller than $Q1 - 1.5 \cdot IQR$ or greater than $Q3 + 1.5 \cdot IQR$. By this point in the conversation, many more voices were being heard. From the video, it becomes apparent that teachers’ conceptions were being challenged and they were being provided with an opportunity to refine their ideas. Within the first two hours of the session, the comfort-level visibly changed and teachers became more willing to engage in collegial conversation and to confront their own conceptions. Teachers became aware that when comparing distributions, a comparison of IQRs may be inconsistent with a comparison of ranges (Site 3, Tape 2, 2:00). That is, comparing different measures of variability may yield different results (e.g., $IQR_1 > IQR_2$ but $range_1 < range_2$) and require careful interpretation.

Pushing teachers to take a stance and provide evidence for statements. Hoping to evoke additional conflict and argumentation, using part g of the activity, teachers were asked to individually rank the students’ overall performance based on students’ quiz scores from 1 to 5 (Site 3, Tape 2, 4:30).

SM: *Have you had enough time to do your rankings? [long pause] What do you usually do in this situation?*

David: *When teachers calculate grades, they almost always use the mean.*

Teachers acknowledged that they would like to compute the mean; however, time was not permitted for them to do so. Jennifer volunteered and ranked the students Susan, Maria, Tran, Gia, and Jack. Six other teachers agreed with Jennifer. Individual teachers began to argue about their rankings and this arguing became quite passionate. Every teacher in each group communicated that they use the “mean” to compute student grades. This conversation was used to provoke teachers to think about whether and in what ways the distributions of students’ scores represented performance that is the same or different. Many arguments surfaced during this debriefing and individual reasoning was shared for the group. This activity may have been especially important to the negotiation of the social norm of providing justification for statements. It was also another opportunity for teachers to acknowledge lack of familiarity with this kind of analysis (Site 3, Tape 2, 9:50).

Dillon: *I have never done this, to compare boxplots to the means of the students.*

During the whole group discussion, the language of variability and signal in the data was introduced. Prompted to consider whether the performance of the five students differ significantly from one another, given identical median scores for four of the five students, teachers presented arguments. The conversation was nothing less than riveting as teachers constructed and put forth their belief-laden arguments supporting their personal grading processes. The group discussed differences between practical significance and statistical significance (Site 3, Tape 2, 18:15).

Dillon: *I don’t deal with boxplots and grades, I don’t have a system that allows me to analyze grades in this manner, so that’s why the mean is basically*

the statistic that I use to issue grades, and if it's borderline, then I take in other factors, you know, health of the . . . , classes missed, whatever you include into that process. So, that's why I say this is new to me, one reason that I've never done it is I don't have access to it. Um and to do it by hand for a student, I don't have time to do it.

SM: [every teacher acknowledges calculating means when assigning student grades] *That seems to me to carry the belief with it that you, somewhere in your belief structure, you believe that, that average, adding up and dividing by whatever is a good signal, is a good center, is a good measure, that does a good job of summarizing all this stuff.*

By the conclusion of this activity, nearly every teacher had participated in the discussion, many different views had been expressed, a common understanding of how to generate a boxplot had been developed, disagreements had been allowed to incubate, and the goal of achieving “right answers” had been averted. Importantly, the deterministic nature of many teachers’ grading practices had been gingerly called into question and the stochastic nature of interpreting data in the presence of variability had begun.

Connecting representations and contexts and sharing authority for knowledge.

Activity 1.3, Matching Variables to Plots was then introduced to support teachers’ reasoning about distributions using shape and context (Site 3, Tape 2, 25:00). Shape of a distribution is a concept that was notably absent from the pre-assessments and pre-interviews. Only one of nine teachers used shape as a characteristic of distributions when reasoning during the pre-interview. During Activity 1.3, teachers worked in groups to match five given contextual variables to five histogram representations. This was the case in each of the three sites. The facilitator began the debriefing by saying (Site 3, Tape 2, 31:19):

SM: *Talk to me about how did you approach this? And I'd really like to hear from some people that I haven't heard from yet.*

There was about 30 seconds of silence and then with a little more prompting teachers began to volunteer. The teachers presented arguments for their selections and seemed confident in three of their five choices.

There were two variables and two histograms that caused conflict for teachers. Two of the histograms were strongly skewed to the right, one histogram contained gaps between interval categories and the other appeared to have no gaps. The variables in question were (1) number of menstrual cycles required to achieve pregnancy for a sample of women who attempted to get pregnant, and (2) number of medals won by countries in the 1992 Winter Olympics. This is the activity during the session in which teachers voiced their unfamiliarity with shapes of distributions. In particular, very few teachers knew what skewed to the right or skewed to the left meant. The evidence from the video corroborated that from the pre-assessment and pre-interviews (see Chapter V). At Site 1, one teacher became quite insistent upon the fact that he had been teaching students the direction of the skewness and that the facilitator must be wrong (Site 1, Tape 2, 2:31).

Drew: *I'm gonna go home and check my book, but I don't think that's right.*
SM: *Go ahead and look. This one's skewed to the right [with a big smile].*
Drew: *Maybe I've just been telling them wrong.*

This signaled that for this teacher, the textbook may have been the mathematical or statistical authority. He did later confirm that he checked the book and had been teaching this and that he interpreted the textbook differently from what was suggested in the session. This was also the activity that challenged teachers' ideas of the difference between bar graphs and histograms. In each of the three sites, there was an example of someone who was clearly thinking about the number of medals won in the 1992 Winter Olympics as a bar graph using country as a category. This confusion provided a space for

clarification between the two representations and appeared important for the group's understanding.

This problem also generated at least one argument from each site which may have positively impacted the ultimate culture of the group. Because the context of one of the distributions was “number of months to achieve pregnancy,” many teachers seemed to argue from a personal perspective. For example, while making an argument in favor of one of two right-skewed distributions, one teacher put forth the following explanation (Site 1, Tape 1, 1:58:00):

Elliot: *Because if they're trying to get pregnant, then* [pause]
SM: *If they're trying to get pregnant?*
Elliot: *Yeah, they will probably be able to do it in one cycle.*
SM: *In one cycle?*
Classroom: [lots of immediate feedback and laughter]
Elliot: *I'm single OK, I don't . . .*
Classroom: [lots more laughter]

This context appeared to be one that teachers could relate to and at the same time, was provocative for their reasoning, sense-making, and communication. The laughter in this situation was seen as evidence of the group bonding in a respectful way. It did not appear that teachers were laughing at their peer, but rather laughing with him. Both humor and vulnerability were supported during this and future episodes.

What happened during this activity may have been significant to teachers' understanding of *comparing distributions* for at least the following four reasons: (1) the discourse during the debriefing allowed teachers to clarify their understanding of shapes of distributions; in particular, they determined the differences between unimodal, bimodal, skewed left and skewed right distributions and associated shapes of distributions with contextual information; (2) this activity initiated a discussion that surfaced the

notion that some teachers were not sure about what a histogram represented; there was discussion of the difference between a histogram and a bar graph that appeared productive; (3) teachers related shapes of histograms to shapes of boxplots; and (4) teachers posed arguments for their choices, articulated their thinking, and vetted their arguments without judgment.

Extending and connecting teachers' reasoning with boxplots and contexts.

Following this activity, a subsequent activity for which teachers were provided sets of side-by-side boxplots representing one of four variables for four different regions of the United States. They were asked to match sets of boxplots to regions of the country and to defend their choices. The activity was selected to further support teachers' boxplot and contextual reasoning and to provide continued pressure to make sense of statistical representations. Teachers' arguments continued to evolve. This was one of many times during the session where even following an extremely rich and evidenced-based conversation, teachers asked, "Did we get the right answer?" (Site 3, Tape 2, 1:10:00). The facilitator took the opportunity to reinforce the norm of making data-based, evidence-based arguments and attempted to convince teachers of the value of their thinking. The attempt was made to help teachers accept some of the authority for the knowledge in this classroom. This activity preceded lunch on the first day of the session.

The next episodes come from Site 2, but at roughly the same place in the activity sequence as the previous Site 1 episode. The episode is selected to illustrate that though the participants were different and the size of the group smaller than that at Site 3, parallels may be drawn in terms of the nature of activities and discourse.

Introducing the use of Fathom2 and exploring measures of variability. Directly following lunch on Day 1, to introduce some of the functions and capabilities available in *Fathom2*, the facilitator collected the number of coins in teachers' pockets and recorded the data in a *Fathom2* collection (Site 2, Tape 2, 16:00). She then talked through the process of investigating the collection, making graphs, generating formulas and plotting measures. The general discourse pattern involved the facilitator asking questions requiring teachers to predict where and how to navigate in *Fathom2* to accomplish the task at hand. This act of requesting teachers to anticipate and predict moves in the technological environment may have been instrumental to their initial learning of the software. Also, the interrogation of the representations by teachers suggested an active stance toward trying to understand statistical representations, measures, and the use of technology.

The next activity illustrated a continued apprenticeship into navigating in *Fathom2*, a prelude to natural variability and making comparisons between distributions, and a launch for the upcoming *Patterns in Data*, Course 1, Unit 2, Measuring Variability: The Standard Deviation, Core-Plus Mathematics Project 2nd Edition draft materials, pages 45–74 (see Hirsch et al., 2008). Teachers were asked to measure their handspans. A sample of the discourse is presented below (Site 2, Tape 2, 23:40).

- SM: *You have rulers at your desks? OK, can I have everybody measure the length of your handspan . . . so as wide as you can.*
- Teachers: *In what units?*
- SM: *Oh, what would be good units?*
- T1: *Centimeters.*
- SM: *OK, how about we go with centimeters from the end of the tip of your thumb to the end of the other side of your pinky finger. So come up with your measurements . . . Got them? So everyone in your group has the same measurement, right?*
- Tracy: *Uh, no.*

SM: *Uh, no?* [laughter from teachers] *Why not?*

Sasha: *Because our hands are all different sizes.*

SM: *Your hands are all different sizes? Hands come in all different shapes and sizes and so it would be strange if you all had the same measurement? There's some natural variability in hand sizes. OK.*

Tracy: *Piano hands or not piano hands.*

SM: *Yeah, well last week we had "man-hands" or "not man-hands" [laughter from teachers], anyhow it was kind of interesting. OK, so now you've got those measurements. I'll tell you what, let's just for grins, let us just . . . can you tell me, I'm going to delete that attribute so I have a different one and I'm going to put a new one in here in—handspan— how's that?*
[Collecting the data in a Fathom2 table: 21.8 (wow, you're accurate), 17.4, 23, 20.4, 19.2, 19.2 (oh, you're the same, wow), 23.4, 24.3, 21]
As a rule of thumb with statistics anytime you have data one of the things you always want to do first is make a picture. As you look at that list of data, can you get a sense of what it's going to look like when I graph it? Or would there be something you would anticipate it looking like? [long pause]

Tracy: *I'd say normal, normalish.*

SM: *You'd say normalish? And why would you say normalish?*

Tracy: *Um, you've got some high values and some low values.*

Brock: *19, 20.*

SM: *19, 20, other people think similarly? Let's take a look [makes a dotplot as in Figure 6a] Is that what you'd think? How useful is that, let's see [changing the representation to a histogram].*

Tracy: *You should change the box length.*

SM: *You want to change the box length. Maybe, now here's kind of an interesting thing. Notice I'm right on the edge of that bar and I can change the bin width. Watch what happens [drags the bin width on the histogram until the distribution is shaped like that in Figure 6b.] Oh what does that look like?*

Teachers: *Symmetric.*

SM: *It's, symmetrical, what else? Is it reminiscent of anything else we saw today?*

Teachers: *The heights.*

SM: *Those height ones.*

Teachers: [affirming]

SM: *What'd we call that? Oh I don't think we called it anything, did we?*

Teachers: *No.*

SM: *What could we call it?*

Teacher: *Hilly.*

Teacher: *Bimodal.*

SM: *Bimodal, two modes [gesturing with hands to indicate two peaks], hilly would be another way . . . another type of a thing. So does that look . . . Could we explain that [bimodal shape] and if we could what would we*

try to . . . How would we try to do that? What might we think?

Sasha: We talked before with heights of men and women, listening to the men report their values overall they have a larger handsan than women.

SM: Yeah? So we don't know who our men and women in our, according to our data right now. Could you go back and tell me, what if I put in gender [a new Fathom2 attribute]? Can we go back around. . . .
[Collecting gender data: M, F, M, F, F, F, M, M, F]
One of the things you can do in Fathom2 that's kind of cool is, I can make a graph down here [dragging a new graph onto the screen] of gender. Now gender, of course, is not numerical data, right? So what kind of graph's it going to make?

Teachers: Bar? [not confident]

SM: It's going to make a bar graph. Well, you want to see the females?
[clicking on the bar graph female attribute as in Figure 6b] Is that what you thought?

Teacher: Huh [chuckle].

Teacher: Yeah.

SM: You see what it [Fathom2] can do?

Jaylee: That's neat.

Jordan: What if they were mixed, what would it have done? Would it have been striped?

SM: If it were mixed?

Teacher: Yeah, if they overlapped.

Teacher2: If there was a category where both male and female were in.

SM: So like if I, what is the smallest, if I made this an M right here?

Teacher: Yeah.

SM: OK [making the change], my graph [bar graph] should update, right? You saw it change down at the bottom.

Teachers: Yeah.

SM: Now see what happens when I click the females.

Sasha: Can you change one of the 19.2s to a male because then there would be a male and a female in that category?

Jaylee: [inaudible]

SM: Oh yeah, OK. Now you see this change [the bar graph], right?

Teachers: Oh . . . ahhh. OK.

SM: You see one of each.

Teachers: Oh. It's the frequency. . . .

SM: OK, does this make you start thinking about the ways you can start to investigate things when you're looking for how things are related? So kind of keep that stuff in mind as we go. OK, now um, another thing I could do if I wanted to, but help me go back and fix these [gender attributes] . . . were those the only ones I messed with?

Teachers: Uh huh.

SM: If I wanted to look at this in two different ways, I could grab gender and bring it over here [to the vertical axis on the histogram] and it would split it [see Figure 6c].

Teachers: *Oh.*
 SM: *So if I want to look at them side by side, or if I want to look at two boxplots side by side [see Figure 6d], and start to do some comparison, I can begin to do some of those kinds of things. At just the click of a button.*



Figure 6. Graphical representations from the handspan investigation.

This episode is a brief six-and-a-half minute segment of the first afternoon of the session which illustrates a number of intentional aspects of the professional development design that can be seen throughout. For example, using the handspan data from teachers afforded opportunities to discuss issues of measurement and natural variability with real data. It provided a context in which the introduction to using technology to explore data could be modeled in an engaging and conversational way. The context would both connect back to the previous investigation, Activity 1.3, when teachers argued for

matching variables to plots and forward to investigating measures of variation such as mean absolute deviation from the median and from the mean and standard deviation in Activity 1.4. Teachers' ideas and conjectures were valued and shared in the public space.

The episode represents a modification from Session 1 because the decision was made to introduce teachers to navigating in *Fathom2* while exploring properties of the handspan data, rather than waiting until later in the session. One advantage that can be seen due to this modification was through teachers clearly wrestling with understanding linked representations in *Fathom2*. Almost immediately when the bar graph and the histogram image were dynamically linked, teachers began to ask "what if" questions and the group explored some of these questions with the technology. This is taken to be an indication that the technology was influencing teachers' interpretation of multiple-linked representations and perhaps helping them to discriminate between categorical and numerical data.

Though it is not clear from the transcript, nearly all teachers during this episode appeared engaged in this activity either verbally or through non-verbal body language. For example during the review of the video, there are times in which teachers appeared to be conferring with one another but their voices were not picked up on the microphone and consequently no transcript was possible. Other frequently seen gestures resemble teachers appearing to say, "Oh" and nodding positively, indicating perhaps some type of resolution or understanding. Other instances involve teachers physically motioning some type of agreement or confusion. One of the limitations of the video transcript in this study is due to the facilitator wearing the microphone, thus her voice often dominates the video and frequently other utterances were not directly picked up for lack of volume and

microphone sensitivity. Even with that limitation, a good share of the verbal and non-verbal communication was captured.

The following episode refers to the follow-up activity, but at Site 3. As will be discussed shortly, the larger group size afforded a slightly different activity structure. Slightly more than three hours into the professional development experience (Site 3, Tape 2, 1:12:00), using the handspan data generated previously, teachers explored additional measures of variability. Table group sizes were Group 1: 4, Group 2: 4, Group 3: 5, and Group 4: 6. Groups' measures were ordered by magnitude and recorded for class display. Teachers explored relationships among the multiple measures. In particular, they calculated the mean absolute deviation from the median (Group 1: 1, Group 2: 1.175, Group 3: 1.45, Group 4: 0.72) and the mean absolute deviation from the mean (Group 1: 1.22, Group 2: 1.36, Group 3: 1.5, Group 4: 0.81). During the discussion, the list features of the TI-84 graphing calculator were demonstrated in order to support the calculation of the mean absolute deviation of the mean after it was discovered that very few teachers utilized this capacity of the tool.

A comparison between the groups' measures prompted additional conversation (Site 3, Tape 2, 1:32:12):

- Stephan: *How come the means are always bigger than the medians?*
SM: *How come the means are always bigger than the medians? What makes you say the means are always bigger than the medians?*
Stephan: *Because every one went up and not one went down.*
SM: *Because every one went up, none went down. So the average distance from the mean is higher than the average distance from the median.*
Stephan: *Or, wait, wait a minute. I've got to think about this now . . . They're farther from the mean than they are from the median.*
Dillon: *That's due to the variation in the data.*
Jennifer: *I think it's the mean takes every, the value of every number into account.*
Wilson: *Would the median be a better measure of the data if it's smaller?*
David: *The mean swings more toward the extreme observations.*

SM: *The mean swings more toward the extremes?*
David: *Is affected more by the extremes.*
SM: *OK, so, does that, what does that seem to indicate here?*
Stephan: *So, if you've got a lot more data, then the mean and the median ought to be about the same? Is that what you're saying [referring to David]?*
Scott: *If it's skewed, the distribution [inaudible] . . .*
David: *If it's skewed, the median doesn't respond to that skewness but the mean does.*
Megan: *Well, just looking at our data.*
SM: *Yeah.*
Megan: *The bottom, there were five of us, the bottom two were much closer to the median than the top two were.*
SM: *The bottom two were closer to the median than the top two were?*
Megan: *That would make our average a little higher than the median.*
SM: *OK.*

The whole group was wrestling with making sense of the mean absolute deviation from the median and the mean absolute deviation from the mean. There was no problem with the procedure of determining the measure, but its interpretation was still fuzzy. At this point in the conversation, the facilitator made the decision to input the data from Megan's group into *Fathom2* in order to (1) introduce teachers to the ease with which one can explore data and navigate in the environment, and (2) to follow up on the relationship between the mean and the median in this particular case and then to illustrate the dynamic nature of the drag and compare aspects of the software. The episode suggests that the context of the task and the two similar but subtly different measures of variability provided a venue from which teachers could extend their current understanding. It also provides evidence of teachers beginning to question each other and not just interact with the facilitator.

Of special interest were the ways in which teachers reasoned about which group they expected to correspond with the largest or smallest measure and whether they could identify the groups by their measures of variability alone. Teachers generally determined

that groups made up of similar size people would tend to have the smallest measures, whereas, groups composed of relatively tall males and shorter females may tend to have the largest measures. Because of the variation in group sizes, the effect of sample size on resultant measures could be explored. A productive feature of this discussion was that teachers could make and test conjectures. The group toggled between looking at measures, looking at groups' physical handspans, and reasoning about multiple measures. They could then refine their emerging theory of the meaning of some of these measures of variability in the context of actual body measurements and group characteristics.

This example is also an illustration of a difference between the enactment of curriculum across two groups. Because Site 2 contained only 9 teachers, the potential to explore groups' measures of variability did not allow for the same rich discussion in the larger Site 3 with 16 teachers. With the larger group, measures from subgroups of size 4, 5, and 6 could be compared. Though this difference across sites was unavoidable, the size difference did afford slightly different access to statistical ideas during this particular investigation.

The last part of the activity related the standard deviation to the mean absolute deviation from the mean and introduced its relationship to the normal curve using a handout from the packet (H-22, Day 1). Groups calculated standard deviations for their handspan data (Group 1: 1.67, Group 2: 1.62, Group 3: 1.82, Group 4: 0.93) and compared the new measures with previous measures. Two things that teachers articulated was their acknowledgement that all of these measures were just different measures of variability. Furthermore, a number of teachers asked about the difference between sample

standard deviation and population standard deviation. An upcoming exploration was designed to address that very issue.

Reflecting on first day similarities and differences across professional development sites. Teachers' first written reflection occurred following the completion of Activity 1.4. Sites 1 and 2 completed this activity by the end of Day 1, whereas Site 3 required about 2 hours from Day 2 to get to a similar point. Site 3 was slightly behind the schedule of Sites 1 and 2 on Day 1, due largely to moderately more-extended debriefing discussions and the intentional modeling and navigation of several more *Fathom2* features than in previous sessions. The discussion of the student quiz score ranking and issues of grading was much more extended at Site 3 than the others. The exploration of the handspan data with the *Fathom2* connections and multiple measures of variability and comparisons among groups was more extended than in previous sessions. It was not an intentional veer away from what had been done previously, but rather a function of teachers' engagement with the tasks and their questions and argumentation that promoted the change. Additionally, the facilitator viewed the extended exchanges as a further opportunity to advance positive social and sociomathematical norms for the group. The seeds of statistical reasoning and technology modeling that emerged during the discussions would allow a shortened launch into subsequent investigations.

By the end of Day 1 at Site 3, teachers completed Investigation 4, #1, 3, 4, 5, 8, 9, 10, 11, 12 from CPMP *Patterns in Data*. The need to reconcile the original definition of standard deviation with the Core-Plus Mathematics definition (dividing by n versus dividing by $n-1$) had been generated. Properties of mean absolute deviation from the median, mean absolute deviation from the mean, standard deviation, range, sum of

absolute deviations, and normal distributions had begun to be investigated. As will be evidenced through analysis of teacher written reflections in Chapter V, by the end of Activity 1.4, teachers were acknowledging a greater understanding of a number of statistical ideas that may be traced back to the investigations from the first three activities. Teachers identified issues of standard deviation and sample standard deviation most frequently when asked about the statistical ideas they better understood by the end of Activity 1.4; however, collectively a large number different concepts were mentioned by teachers.

Day 2

Comparing population standard deviation to sample standard deviation. An important activity that occurred on Day 1 at Sites 1 and 2 was delayed until Day 2 at Site 3. The activity was part of Activity 1.4 and involved the investigation of the standard deviation of a sample. Teachers explored sampling of size two from a population containing exactly three elements. Teachers generated all samples, with replacement, of size two, computed the sample means and then two possible sample variances. For one computation they divided the sum of the squared deviations from the mean by n (in this case 2) and for the other they divided by $n-1$ (in this case 1). When they compared the average variance from the two options to the variance of the population, they determined that the second option did a better job, on average, of estimating the true population variance. This led to a discussion of unbiased estimators and the properties of good estimators. It should be noted that the tone of the discussion was more conceptual than theoretical. When discussing estimators that were unbiased, consistent, relatively

efficient, and sufficient, the class drew on estimators that had been explored to that point and made connections to these ideas. Teachers began to sense that different estimators had differing potential to do a good job and that choice of estimators mattered, given the conditions of the context. It appeared that this language was entirely new to teachers. Though this content was not approached theoretically, teachers appeared to grasp the general ideas. The idea of an unbiased estimator seemed to resonate with teachers and repeatedly surfaced during remaining investigations.

Connecting measures to representations. The final portion of Activity 1.4 involved teachers matching statistics to plots. Teachers were asked to match six histograms to their corresponding summary statistics. Summary statistics included mean, median, and standard deviation. This activity supported making statistical arguments, clarified relationships between different statistical measures under varying circumstances, and connected shape of distribution to properties of measures. The second part of the task required teachers to match distributions presented as histograms to those presented as boxplots. Collectively, this culminating activity was rich with statistical connections and furthered the goal of creating a learning environment in which making and justifying statements with evidence was valued.

Introducing teachers to the randomization test. After teachers had been introduced to and wrestled with issues of standard deviation of a population and a sample, the randomization test procedure was developed using the experimental data from the Orbital Express Activity 1.2 (Site 3, Tape 4, 20:30). Teachers physically simulated a trial of the randomization test with one group's data using index cards with

data values on them. They were asked to put the two data sets together, representing the assumption that there was no difference between the vehicles and then to shuffle the cards and redistribute them into two piles, each with seven data values. They were to make a quick sketch of each distribution and calculate the mean of each (Site 3, Tape 4, 28:40). This took more than seven minutes to accomplish, suggesting some lack of understanding of the initial task. *Fathom2* was used to collect teachers' experimental mean differences (Site 3, Tape 4, 38:25).

SM: *You guys are going to get your hands on Fathom2 here shortly, so, just again, to kind of get you in that spirit of things, if I want to put this data into Fathom, I need a collection. I can either make a collection or I can even just grab a table, because as soon as I call this, let's call this uhm*

...

Teachers: *Group 1?*

SM: *What?*

Teachers: *Don't you have group 1 and group 2?*

Teacher: *So did you grab a table?*

SM: *I just grabbed a table, tossed it in there. For the moment I'm going to call these differences, right? Differences in the mean, but I'm going to call them differences. Can you guys read those off to me and I'll put them in here?*

Callie: *1.5, 7.8, 12.9, 7.5, 4.3, 1.1, 3.4, 1.5, 10.9, 10.7, 6.4, 9.4, 2.1, 2.5, 8.4, 4.7, 2.3, 8.5.*

The group proceeded to have a conversation about why all of the values were positive and what can be determined from this new distribution of differences in means. Once teachers acknowledged they reported the absolute value of the difference in means, which the facilitator already had deduced, the data values were revised appropriately and the discussion continued with the building and interpretation of the randomization distribution in *Fathom2* (Site 3, Tape 4, 48:10).

SM: *So I'm building this distribution right now . . . What do you notice about this distribution?* [long pause]

Callie: *Most of it's less than what we had.*

SM: *Most of it's less than what we had in our true experiment, OK, what else*

do you notice?

Dillon: *Only a few values exceed the difference in means of the two papers.*

SM: *Only a few values, and in fact, how many does it look like?*

Teachers: *Three maybe, five, four . . .*

SM: *Looks a little hard to tell here, right? Looks like there are five of them. So five out of how many?*

Teachers: *36.*

SM: *Five out of 36. Would we consider that to be unusual, or could we get results like that by chance alone?*

Dillon: *Statistically, isn't there a number, a p-value for like confidence interval or something for when it falls in or out of expectations?*

SM: *Maybe.*

Stephan: *It's like rolling dice, sometimes you get 7, sometimes you get snake-eyes. It's like yeah, you get more 7's than snake-eyes, but if you just do it once.*

SM: *OK, could you begin to, if you had to make a prediction right now, if we kept up this business of shuffling these cards and doing exactly the process that you did, could you predict what you think might happen here? If we kept doing this based on, we've done . . . 36 of these collectively.*

Tyra: *. . . approaching zero.*

SM: *OK, you think the mean is approaching zero. Does anybody else think that?*

Dillon: *The mean? I think so.*

SM: *Anybody else?*

Megan: *I'm not so sure, because I took the mean the first time [with 18 trials], and the second time [with 36 trials], and it's getting closer to 2.*

SM: *The mean is getting closer to 2.*

Megan: *2.02 and now it's 2.016.*

SM: *So it's a little higher than zero at this point.*

In order to develop a better sense of what might happen in the long run, *CPMP-Tools* was used with the original data values to simulate the building of the randomization distribution (Site 3, Tape 4, 53:00). Parallels were drawn by teachers between what *CPMP-Tools* was doing and the physical simulation they had done. Teachers began to ask a lot of questions about different representations on the screen as the simulation ran and the distribution was being built. Once 1,000 trials were run, teachers said the distribution looked “kind of normal,” “centered at 0,” “symmetric” and made a reasonable argument for the why the distribution might be centered at 0. When asked

how they would decide how they might figure out the probability of the mean difference of 11.1 or greater occurring by chance, teachers suggested using the relative frequency from the empirical distribution. From there, the group tried to determine whether that probability suggested a significant difference (Site 3, Tape 4, 1:02:00).

- SM: *So 108 times out of 1000, so about . . .*
Teacher: *11%.*
SM: *About 11% of the time, let's call it about 10% of the time, we're going to see results as big as 11, just by chance alone. Right? If there is no difference between these papers, we are still going to see results that different about 10% of the time. The question is, "Is that unusual?" Would we say there is a significant difference between those vehicles based on what we're seeing here [gesturing toward the randomization distribution on the screen]? [long pause] Jennifer, you're shaking your head no. Why is it no?*
Jennifer: *I don't know.*
SM: *Why are you saying no?*
Jennifer: *I still think that 10% is that often for, to be able to justify there is a difference.*
SM: *OK.*
Teachers: *[Inaudible]*
SM: *I don't know, say it again.*
Jennifer: *I don't know. It just doesn't seem . . .*
Wilson: *Wouldn't the thinking be just the opposite?*
Teachers: *[Lots of inaudible talking]*
SM: *OK, so let's work together here to improve this thought.*
Megan: *My thinking is that 10% of the time, we'll get results like that, so them getting a result like that means that it's not just chance, it's probably something else.*
Wilson: *That's what I was thinking, but I might change my mind in a minute.*
SM: *That's what you were thinking but you might change your mind in a minute. Does somebody else have a thought on this? You want to add something to it?*
Blair: *No, it was pretty much along the same lines as that.*
Dillon: *Cause like I say, I can never keep this stuff straight which is part of the problem, but it seems to me that whether it's significant or not depends on this level of confidence that you want to be able to have. Depending on where you choose that, I thought you had to choose it before you started the problem-solving but in this case, it could be significant or may not, depending on what level of confidence you want to have, and we didn't choose that.*
SM: *OK.*
Dillon: *So that's what keeps confusing me by your question of is it significant or*

not, we haven't established, at what point do we base that decision on?
 SM: *Ahh, OK.*
 Dillon: *That's what I'm bothered by.*
 SM: *So 10% of the time, even if we, even if there was absolutely no difference between these things, we would see a result like this. This unusual, right? The question is, "Is 10% of the time weird enough?"*
 Beatrice: *Could that have something to do with standard deviation? If it's outside of the standard?*
 SM: *OK, so we want to make a . . .*
 Beatrice: *That would be considered significant or not?*
 SM: *OK.*
 Beatrice: *Is that where that comes into play maybe?*
 SM: *Well there certainly can be a relationship with the standard deviation.*
 Beatrice: *One standard deviation is 2/3 of the data, so if it's in the rest of the third, then that would be . . .*
 SM: *Unusual?*
 Beatrice: *Yeah, [inaudible] . . .*
 SM: *So what I think I'm hearing you guys say together is that since it only happens 10% of the time, that you're saying that there is some evidence to suggest that there might be a difference.*

The conversation continued as teachers argued for and against various perspectives. A definition for "significant difference" had not been formally stated. Teachers were drawing on their prior knowledge and trying to create arguments for their perspective. Several minutes later a reference to another of the group's data sets from the Orbital Express Activity was suggested as a contrast (Site 3, Tape 4, 1:08:30).

SM: *Would you say these two distributions [referring to another group's poster] look significantly different? Would you expect their means to be significantly different? Let's see what happens here [loading the data into CPMP-Tools].*
What I want you guys to see, and I want to have a discussion about, is how these things can play out. And looking at those pairs of boxplots [pointing to original data sets] and looking at those pairs of boxplots [pointing to the new comparison], and how this randomization process kind of goes.
 [Group 1: 91.5, 107.3, 45.7, 55.5, 22, 89.5, 99.3
 Group 2: 29.1, 19.6, 16, 26, 29.8, 20.4, 37.1]
The difference in the means is 47 units as opposed to the 11 we had before. I'm going to start this [the randomization distribution in CPMP-Tools] and see if it matches what you think is going to happen. It's

shuffling and computing. You want me to make it go faster? [long pause] Is it behaving in a way you anticipated? So we've done 500 runs, what do you notice?

This whole bar right here [pointing to the most extreme bar on the histogram] contains only eight pieces of data. At this point, what's difficult for me to tell right now?

Callie: *How much is really outside the line.*

After changing the bin widths based on teachers' recommendation, the group determined only 2 of 500 (0.4%) runs produced a difference in means of 47 units or greater. Teachers saw that the results from the second simulation indicated a far more unlikely scenario than that from the first simulation (Site 3, Tape 4, 1:13:00)

SM: *Significant difference?*

Scott: *More so than this one [referring to the first poster].*

SM: *So you're seeing more of a significant difference here [second simulation] than you did here [first simulation]?*

Scott: *Because of what we said yesterday too.*

SM: *Because of what we said yesterday too? What was that?*

Scott: *This one [pointing to the poster used for second simulation], I think this was the one that was the closest yesterday. We argued about this one being significant over all the rest of them.*

This clip provides another example of how previous activities or artifacts may have been invoked to impact teachers' understanding. It also illustrates the sense of "relativeness" that was emerging through various comparisons.

Beginning explorations with Fathom2. To support teachers' use of *Fathom2* to conduct statistical investigations, the next activity introduced teachers to personally using the tool through tours included with the software package and designed to familiarize users with navigating in *Fathom2*. Teachers were instructed to work in pairs to complete *Fathom2* Tours 1 and 2. Tour 6 was recommended for teachers to explore if time permitted. Most teachers completed Tours 1 and 2, suggesting they had some notion of

how to establish a collection, create and manipulate graphical and numerical summaries. Questions about *Fathom2* that had arisen during the tours were addressed and the process of randomization testing reviewed. Teachers then completed Written Reflection #2.

Using Fathom2 to build a randomization test mechanism. Following the reflection, the facilitator modeled building the randomization distribution mechanism in *Fathom2* as teachers connected the physical experience and the actions seen in *CPMP-Tools* to the *Fathom2* motions. Teachers were then asked to take a few moments to write down what they thought they needed to remember from the demonstration so that later they could build the randomization distribution in *Fathom2*. The facilitator had the sense, and the videotape seemed to confirm, that teachers quite readily connected the physical simulation to the *CPMP-Tools* Randomization Distribution simulation to the *Fathom2* environment. As will be discussed in the Day 3 section, teachers were quite successful on their first attempt to generate the randomization distribution mechanism in *Fathom2* with really limited technology exposure.

Introducing sampling ideas and creating disequilibrium. The following activity sequence introduced teachers to sampling, types of sampling bias, and the influence of sample size on sample means. Teachers were asked to use the Random Rectangle sheet to (1) predict the average area of the 100 rectangles on the page, (2) take a sample of size 5 from the rectangles and compute the sample mean, (3) use a random number generating device (random number table and calculator) to select a random sample of size 5 and compute the sample mean, and (4) take a random sample of size 10 and compute the sample means. After the first prediction, the predictions were collected for the whole

class in *Fathom2* and the distribution displayed. Teachers were asked whether they wanted to revise their predictions. Then the self-selected samples of size 5 sample means were collected and the distribution displayed. Predictions were revised. Then random number generating devices were introduced and used to create samples of size 5; sample means were displayed and compared to previous distributions. Finally, random samples of size 10 were collected and displayed and the distribution compared. When asked about which of these methods teachers would choose if they really wanted to predict the mean of the population, the following exchange occurs (Site 3, Tape 5, 39:00):

- Amy: *Do larger random samples.*
 SM: *Do larger random samples? Why? What's the benefit?*
 Beatrice: *Well, your standard deviation is smaller, which means you're getting closer to a constant value.*
 Rianna: *It's also easier to calculate the average. I came up with 5.4. 54 divided by 10, you don't have to do . . . it's a mental math thing.*
 Celeste: *Isn't the lower your standard deviation, the more valid and reliable your data is?*
 SM: *Group, what do you think about that? She's saying that the lower the standard deviation, say it again.*
 Celeste: *The lower your standard deviation, the more reliable, is it? More valid and reliable?*
 Beatrice: *In this case [sample of size 10], what we're looking for is 7.42 and our standard deviation is lower.*
 Amy: *A lower standard deviation means your data are clustered, doesn't it?*

The conversation at this point in the session indicated that teachers were really attempting to coordinate the many ways of thinking about variability they knew. It also represents the whole group continuing to develop a shared understanding of the meaning of standard deviation. At the conclusion of this activity, teachers were asked to write down what they took away from it. As a pedagogical move, the facilitator was encouraging teachers to reflect on the experience to clarify and explicitly attend to their own thinking. Throughout the continuing conversation, issues of bias, center, standard

deviation, range, sample size, randomness, sampling variability were all addressed.

Written Reflection # 3 which occurred later in the session (results reported in Chapter V) provides some of teachers' responses to these issues.

The follow-up portion of the activity was designed to elicit shock and awe (Site 3, Tape 5, 48:00). Essentially, a video clip was used for which teachers were instructed to count the number of times basketballs were tossed among a group of students in a gymnasium. The clip was set up as a data collecting task. After watching the video, teachers' estimates were verbally solicited around the room; estimates exhibited much variation. The facilitator then asked how many people saw the gorilla. About 1/3 of teachers acknowledged seeing the gorilla (less than 1/3 of teachers in previous sessions). The remaining teachers began to laugh and question whether there could have been a gorilla in the video. The group watched the video again and much laughter erupted. Teachers were amazed that they missed the gorilla the first time around. When asked why the facilitator chose to use this video clip, a number of responses suggested that the point of potentially missing the forest for the trees had been accomplished. Some teacher reactions included: "What things get by when you're concentrating on the data," "You see what you're looking for, and not necessarily other things," "In trials, eyewitnesses may not be very reliable," "Kids don't always see the same things that we want them to see," and "How much do you trust your numbers if you don't see a gorilla walking across in front of you?" Overall, this activity was pretty light-hearted yet powerful for teachers and facilitator to continue to build a safe environment in which ideas were valued, determinism was challenged, and sensitivity to data collection and sampling were reinforced.

The final part of Day 2 involved teachers beginning to explore experimental design and conditions for establishing cause and effect with Core-Plus Mathematics, Course 3, Unit 1: *Reasoning and Proof*, Lesson 4, Investigations 1 and 2. Teachers were responsible for completing Investigation 1 and from Investigation 2, problems 3 and 4 prior to Day 3. The debriefing of these investigations would take place on Day 3.

Day 3

Exploring experimental design and reasoning about causation and statistically significant differences. Day 3 began with a debriefing conversation of the Core-Plus Mathematics investigation that was assigned for homework. During the debriefing of Investigation 1, teachers clarified the meaning of a placebo, lurking variables, the need for blind and double-blind studies. Further, issues of random assignment, sufficient number of subjects, and control were explored. Investigation 2 then was used to reinforce the connection between experimental design and the interpretation of the randomization test. Questions 3 and 4 provided a context and example of an experiment in which the difference in the means of two groups of experimental data were not significantly different. The follow-up, Question 6, then provided a context and an example of an experiment in which the difference in the means of two groups of experimental data were highly significant. Teachers were asked to make predictions based upon the dotplot or boxplot representations and then to interpret the randomization distribution of the differences in means in both cases. The combination of these three problems with the previous Orbital Express investigation produced something of a syzygy of sense-making as evidenced in the following exchange (Site 2, Tape 5, 1:24:00):

Jordan: *There seems to be a difference there.*

SM: *You're looking at dotplots right? I don't think I can get a dotplot here [in CPMP-Tools], but I can get a histogram. Same data. About like so? [teachers nod in agreement]*

Tracy: *Can you do a box and whisker for that one?*

SM: *OK, so say why.*

Tracy: *Because I want to remember what the box and whisker looked like for scented and unscented.*

SM: *OK.*

Tracy: *That was one where it wasn't significantly different.*

SM: *Yeah.*

Tracy: *And so I remember how those looked [from problems 3 & 4] and so I'd like to see how those [from problem 6] look.*

SM: *Alright, excellent. Everybody have this kind of picture? OK, you have it another way. Let's look at the boxplots. Having the flexibility to look at these things in different representations might be helpful.*

Tracy: *It does look kind of different.*

SM: *So what do you think there [referring to the boxplots on the overhead screen]? That looks a little different than the other one, doesn't it?*

Jacob: *Yeah, it looks more kind of like the Orbital Express, where you had, there wasn't really a lot of overlap with the mass, they're pretty much overlapped one another, where with that experiment there, they're kind of offset.*

SM: *So the boxes here are . . .*

Tracy: *More shifted.*

SM: *There's no overlap.*

Tracy: *Like look at their last Orbital Express one [referring to a poster on the side wall].*

SM: *Yes.*

Tracy: *Now looking back at that, I might look at that and say I'm not really confident that there's a big difference between those two.*

SM: *OK.*

Tracy: *Looking at the means being so close together.*

SM: *Are you looking at the means or the medians? They've got them both there.*

Tracy: *Medians, but the means are too.*

SM: *OK.*

Tracy: *Like our box and whisker plots, I would feel confident saying that they're pretty different.*

SM: *You'd say they're different?*

Jaylee: *The packages were different when we did actually do the randomization.*

Tracy: *They were?*

Jaylee: *Um hum.*

Jordan: *They were like 7%.*

Tracy: *Oh yeah, the girls had a higher . . .*

Jaylee: *Um hum.*

SM: *This one here?*
 Teachers: *Um hum.*
 SM: *These weren't significantly different [referring to one poster]; those were significantly different [referring to another poster]?*
 Teachers: *Um hum.*
 Tracy: *But do you, do you girls, I'm not allowed to call you girls, "women" [forcefully but with apparent affection] . . .*
 All: *[Laughter]*
 Tracy: *Do you remember when you first did that? Like the very first day, did you look at your plot and say, "Oh, I think there is a difference?"*
 Jaylee: *Yeah.*
 SM: *Everybody said that.*
 Tracy: *Everybody did?*
 SM: *Everybody said that. So now you're thinking maybe . . .*
 Tracy: *I'm all skeptical. Thanks, Sandy [laughter].*
 SM: *Sorry, it's my job [more laughter]. OK, we have some reason to think that maybe these boxplots are helping us to determine this. So now your conjecture was, "Maybe they are significantly different?" [referring to problem 6 again].*

This roughly six-minute clip illustrates a number of things from an analytical point of view. First, the choice of problems with which to engage teachers was essential for the conversation to emerge as it did. The contrast between the context of problems 3 and 4 versus that of problem 6 afforded an immediate comparison between data and representations, one in which significant differences between the means of two groups did not exist and then one in which significant difference in means did exist. As seen in the previous dialogue, teachers were making comparisons by asking for additional representations and beginning to conjecture about relationships. The use of *CPMP-Tools* to quickly change representations and to conduct the randomization test in real time allowed at least two important things to happen: (1) the boxplot representations that were not printed in the text materials became available, and (2) the randomization distribution that was printed in the text could be compared with the simulation being generated and displayed on the overhead screen. The manipulation of the bin sizes through the

simulation allowed the variability of the shape of the distribution to be explored and the idea of p -value emerged at this point in the discussion. At 500 trials, the simulation had failed to generate a difference in means as large as the experimental difference, suggesting an approximate p -value of 0. Teachers discussed the fact that the p -value would not be 0, but rather something positive because eventually the simulation would produce two groups that were composed of the same values as the original experiment. This conversation may have been pivotal to teachers' understanding of significant differences. Secondly, as teachers were presenting arguments, they engaged the Orbital Express activity from Day 1. The posters from that activity were artifacts on the side wall. The presence of those posters as well as the arguments from the first day, appeared to be invoked during this Day 3 conversation to support teachers' sense-making.

Following this discussion, the group explored the connection of bias and variability with that of accuracy and precision using the representation of a dartboard. The conversation allowed multiple views to be expressed and for the group to further refine their ideas of bias and variability. Nearly every voice was heard during this discussion and teachers who had been less verbally engaged up to this point, were able to put forth important claims with evidence (Site 2, Tape 5, 1:52:00).

Reinforcing and extending the use of Fathom2 for conducting the randomization test. The group then moved to the computer environment to explore Activity 3.2, The Seattle Real Estate Task. During this investigation, teachers were able to verbalize the procedure and construct the mechanism in *Fathom2* to do the randomization test. Teachers discovered that this context resulted in contradictory comparisons when different measures were used. The differences in mean selling prices for homes from

2002 to 2001 was \$40,000 and not significant; the differences in the median house prices was -\$15,875 and not significant. That the magnitudes of these differences were so large and not significant was a cause for disequilibrium for teachers. Additionally, since the mean and median differences in home prices were opposite signs, teachers were again encouraged to consider the importance of choice of measures.

Exploring the normal distribution and its connection to standard deviation. To support teachers continued reasoning, Activity 3.5 allowed teachers to explore Core-Plus Mathematics Course 3, Unit 3: *Samples and Variation*, Investigations 1 and 2. To launch the investigation, teachers were provided with a certificate to recognize their hard work on the randomization test and use of *Fathom2*. They were asked to individually determine the perimeter of the certificate border and then individuals' data were collected and used to illustrate measurement variability. Teachers worked in groups to complete Investigation 1, numbers 1-6 and Investigation 2, numbers 2 and 4. This activity was included to support further understanding of normal distributions. Teachers completed the pieces of the assignment that they were unable to complete during the session as homework.

Day 4

Confirming teachers' understanding of normal distributions and standard deviation and supporting mathematical connections. Debriefing Activity 3.5 began the final day of professional development. Through the conversation, teachers communicated a fairly strong understanding of the mathematics of the investigations and appear confident in their understanding of standard deviation in relation to normal distributions.

During the debriefing discussion of the normal density curve from Activity 3.5, the facilitator posed the questions, “What is π ?” and “What is e ?” (Site3, Tape 7, 1:11:45). Most teachers, although not all, were able to supply an explanation of the ratio of circumference to diameter for a definition of π ; however, not one teacher could say what e was. One volunteered, “I can remember saying what e is, but it’s gone now.” Another said, “Isn’t it the sum of a series?” (Site 3, Tape 7, 1:15:49). The group briefly discussed a mathematical limit definition of e followed by a teacher demonstrating the function/table connection with the graphing calculator. The group then explored the normal density function, $y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, via formula and *Fathom2*. Using sliders for mean and standard deviation, teachers made the connection between changing the mean and generating a translation. Similarly, the connection was made between changing the standard deviation and stretching or shrinking (Figure 7).

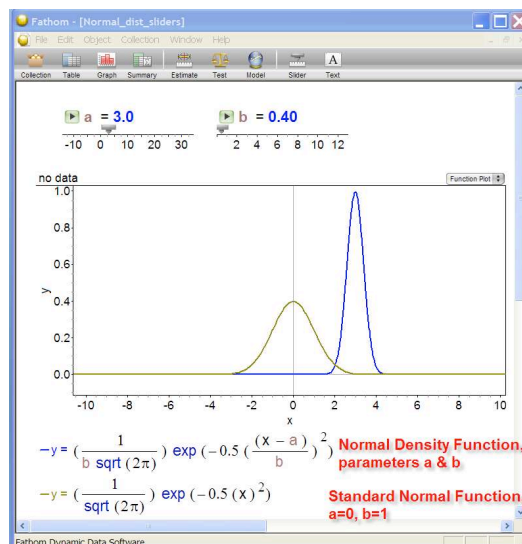


Figure 7. Sliders in *Fathom2* to dynamically model parameter changes to the standard normal density function.

The dynamic use of the sliders appeared to help teachers connect the function representation to the graphical representation and to perhaps further their understanding of comparing normal distributions. Using a demonstration with plotting linear functions of the type $y = mx + b$, it appeared that teachers were considering the use of this tool for potentially supporting the teaching of algebra. Further evidence for this came from teachers continuing to work on and ask for guidance regarding function plotting during the break (Site 3, Tape 7, 1:50:00). This technology demonstration with group participation is an example of foreshadowing a capability of *Fathom2* that may be useful to teachers for an activity that will come later in the session. It also represents an activity with potential to influence or reinforce other mathematical connections for teachers.

During the next scheduled break, Jennifer said, comparing the use of the graphing calculator to *Fathom2*, “this [*Fathom2*] is much niftier . . . because I have precalc . . . I think it would help me. The first time I taught it, I didn’t completely . . . understand” (Site 3, Tape 7, 1:50:40). She was using the break time, as were other teachers, to explore the use of sliders for parameters on function plots. This clip signifies teachers connecting this tool’s use to their classroom practice, continued teacher vulnerability, and curiosity about the capability and value of the tool.

Challenging teachers’ use of Fathom2, their understanding of sampling, and discovering the Central Limit Theorem. After the break, Activity 3.4 began. The activity was referred to as the “crying activity,” due to the fact that during the same activity at Site 1, a great deal of frustration with the use of technology was evident and one very experienced and competent teacher ended up in tears. The limit of educative discomfort had been exceeded. That was a big signal about the level of scaffolding that may have

been beneficial in this activity that was lacking in the first round. Clearly, the first time with this activity, the facilitator made some assumptions about teachers' ability to easily transfer the use of the tool in one context to that of another. Subsequently, minor modification to the setup of the task prevented the level of discomfort from becoming debilitating (Frykholm, 2004). As a prelude to the activity, the facilitator acknowledged the potential challenges of the upcoming activity and also took a moment to thank teachers for their written reflections and make a suggestion that written reflections are useful tools for use with students. The transcript of the beginning of the activity is below (Site 3, Tape 7, 1:53:56).

- SM: *OK, this is the crying activity [laughter]. So we're not going to have any crying today [more laughter]. There's no crying today, we'll work it out.*
- BL: *Do we have Kleenexes?*
- SM: *So I'm forewarning you, but I'm going to tell you that even if you get frustrated and you feel like this, there's help. Don't get so frustrated that you want to cry, just ask for help, OK? We'll get you.*
- Scott: *Boo hoo.*
- SM: *Epecially you, Scott [laughter].*
OK, so I want to bring you back. We've been doing all of these kinds of things and we've got all these big ideas from statistics floating out there. I think for the next couple of minutes, what I need people to do is just flip those laptops right down [more laughter]. Just flip 'em down.
- Teacher: *See Beatrice, you ruined it for everybody. [laughter]*
- SM: *OK, so there's a lot of stuff in statistics that to have a really good idea about some of the things that are going on, some of the things are sort of subtle. And when I was reading your reflection things, and I wanted to say that first of all, I appreciated the fact that you guys took the time to write as much as you did, as I was entering all that stuff into a database last night at about 12:30, I thought it was awfully generous of you to share your thoughts like that—like you did. But it's really helpful for me to hear, from a teaching standpoint, at that moment that you're understanding, and so that I can use that information hopefully in a productive way. So if you don't do that kind of thing in your classroom with your kids periodically, I'd give it a try. I think you would be amazed at the type of feedback it provides for you and what it can do for both your kids and for you as a teacher . . . But anyhow, this language issue of what's a p-value, what does significance mean, what is a sample, what is a population, what is x-bar and mu and sigma and stuff?*

All that stuff is sort of out there initially until we're able to, as a group, or in your classroom, you are able to develop a shared understanding of what these ideas mean.

As the activity begins, the facilitator referred back to Activity 2.4, Random Rectangles, which had been investigated on Day 2. She reminded teachers of the investigation and teachers acknowledged remembering the process of sampling and calculating measures, and collecting measures to create a sampling distribution. She then orients teachers to the use of *Fathom2* with the Random Rectangle document to simulate the generation of a sampling distribution. This mechanism for generating a sampling distribution of sample means is slightly different from that for the randomization distribution and this scaffolding appeared to be helpful. Through the interactive conversation and facilitator modeling the statistical process and use of *Fathom2* with teachers' help, teachers were able to complete Activity 3.4 with no tears. Pairs of teachers investigated populations of their choice and used *Fathom2* to construct sampling distributions of the means for a variety of sample sizes to complete the following two-page investigation (see Figure 8).

Teachers explored populations including random normal populations (normal), random integer populations (uniform), the product of two dice rolls (multi-modal), average homeruns for major league baseball players (skewed right), batting averages (skewed left) murders in Chicago and the ages of the victims (bimodal), average life expectancy of women around the world (skewed left), maximum temperature of Iowa (bimodal), average age of males mid-life crisis (creatively defined—randomnormal [50, 20]), and other populations of their choice. All groups were successful completing the table components of the activity and recognizing that as sample size increased, the range

Is There a Relationship Between the Sample Size and the Sampling Distribution's Standard Deviation?

You will need a partner and a large population of measurement data for this investigation. You may either

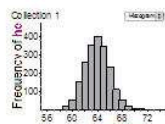
1. open an existing *Fathom2* data collection, OR
2. generate one by creating cases and then using a formula to generate values for each case.

To Generate a Population

Create a collection. With the collection selected, click on **Collection** from the main menu and select **New Cases**. Generate at least 1000 cases (or a proportionally representative model). Suppose that the heights of adult women are normally distributed with mean 64 inches and standard deviation 2 inches. To generate a population of this type in *Fathom2*, double click on the **Collection** to open the **Inspector**. Define a name for an attribute, maybe "height". Double click on the **Formula** column to pull up the formula generator. Using the existing attributes and functions, you can drill down to get the commands needed to generate your function. For this one, double click on **Function**, then **Random Numbers**, then **randomNormal**. You should see the syntax displayed in the editor. You need to provide the parameters 64 and 2. The editor should look like

`height=randomNormal(64,2)`

If successful, a histogram of your data should look similar to the graph below (this graph represents 2000 cases).



You may use this data or generate another set of your own.

You might experiment with other functions available in the Formula Editor in *Fathom2* or try to generate an interesting population on your own.

Use your population to create samples and associated measures that you can collect to complete the following table. In *Fathom2*, you will need to

- 1) Create a collection,
- 2) Sample appropriately from the collection,
- 3) Collect measures from each sample and summarize them accordingly

Sample Size	# Trials	Sampling Distribution Mean (of sample means)	Sampling Distribution Standard Deviation (of sample means)	Sampling Distribution Range (of sample means)	Sampling Distribution Shape--Sketch (of sample means)
1	500				
2	500				
4	500				
10	500				
20	500				
50	500				
75	500				
100	500				
200	500				

Describe any patterns that you see in the table.

Create a mathematical model that represents the relationship between sample size and standard deviation of sample means. Be prepared to defend your choice of model and to explain the meanings of the coefficients in your model in the context of this problem. (For example, if your model is linear, the general equation is $y = mx + b$, the m and the b mean something in the context of the problem, but what?)

Figure 8. Sample activity: Is there a relationship between the sample size and the sampling distribution's standard deviation?

of the sampling distribution tended to decrease, the shape of the sampling distribution became unimodal and approximately normally-shaped, and that the standard deviation of the sampling distribution of the mean (i.e., standard error) decreased. In essence, they were poised to discover the Central Limit Theorem.

During the exploration of a mathematical model that might explain the relationship between sample size and standard error, many teachers expressed uncertainty about how to proceed. Importantly, across all three sites, there were numerous teachers who did one of two things: (1) used the graphing calculator to put sample size in List 1, standard error in List 2, make a graph, and then proceeded through the various regression options on the calculator until they found the best choice as determined by the largest correlation coefficient; or (2) communicated that they did not know how to make progress on this. For example (Site 3, Tape 8, 1:09:30):

SM: *Where did this curve come from or what is that?* [pointing to graph on calculator]

Alexandra: *That is our sample sizes and our standard deviations for our . . .*

SM: *How did you get these things to be connected?* [referring to the points on the scatterplot]

Alexandra: *We did a line scatterplot.*

SM: *Oh.*

Teachers: [laughter]

SM: *I was thinking you have a model for that. But that's really what you want. You want an equation that'll generate something that looks about like that.*

Kelsie: *Well, and we think that this program [Fathom2] should do that. We should be able to tell it . . .*

Alexandra: [more laughter] *Click on the line, give me the equation.*

SM: *It can, in a way.*

Kelsie: [inaudible] *We just have to figure out how to do that part.*

SM: *So, in the absence of being able to figure that out.*

Kelsie: *To have the computer do that for us.*

SM: *Let's think mathematically about it.*

Kelsie: *OK* [sigh].

SM: *What kind of function does it appear to be?*

Kelsie: *Well that's where we got off on the exponential . . . but we were not very good at this* [more laughter].

A few minutes later, groups were still working and the facilitator visited the same group again. She suggested they modify their graph so that the points were no longer connected and they would be better able to determine whether a model fit the data well. Some of the teachers' mathematical thinking about modeling and reasoning with functions is revealed in the following transcript (Site 3, Tape 8, 1:12:30).

Alexandra: *I don't know how to come up with that.*

Kelsie: *We're trying to generate . . .*

SM: *So tell me a little bit about what kind of function . . . what kind of functions come to mind when you see graphs like that?*

Alexandra: *I think exponential just because it's starting at one point and then it curves.*

SM: *OK.*

Kelsie: *It looks like it's approaching limits.*

SM: *All right.*

Teachers: [inaudible]

Alexandra: *How you find those, I don't know.*

Kelsie: *It looks like it would be zero and whatever it would be out here, depending on what our sample size was.*

SM: *OK.*

Kelsie: *It's approaching zero.*

SM: *And let's see. What's approaching zero, what's getting smaller and smaller is what value?*

Kelsie: *The standard deviation is approaching zero as our sample size gets higher and higher.*

SM: *Can the standard deviation ever get to zero?*

Kelsie: *No. It can approach it, but it will never get . . .*

SM: *It can approach zero, all right, but we're never going to cross over it?*

Kelsie: *No.*

SM: *OK, that's a good thing. Are there some other functions that behave that way?*

Kelsie: *This is where our knowledge is way limited [laughter]. I blame myself for never teaching Algebra II—for the last 10 years.*

SM: *Actually, I'm hopeful that seventh graders are going to come out with this.*

Kelsie: *Oh my.*

Alexandra: *With this?*

SM: *Well, maybe not with this relationship, but relationships that are similar to this. What do you think about, when something gets big, something else gets little?*

These teachers went on to think about inverse variation and tried to determine an appropriate model for the situation. These excerpts are taken to suggest that teachers felt safe enough to expose their mathematical thinking and potential weak areas to each other and the facilitator. They may have learned by this time in the session that the facilitator would neither tell them what to do nor give them the “right answer.” Though they were struggling with the mathematics required of them, they did manage to build on what they knew and apply their knowledge in a sense-making way. They were not being judged, but they were being challenged. When the facilitator suggested that the mathematics of this task might be reasonable for seventh graders, that is, inverse variation is appropriate for seventh graders to investigate and model, the teachers did not even bristle. It was as though they understood that the group was working together to support the collective

learning of all members and if seventh graders could do this, then so could they. And they did.

As the whole group shared their mathematical models and explanations for how they arrived at those models, the connection between mathematics teaching and statistics learning was utilized again. Teachers were pushed to articulate the meaning of the parameters in their models. Even the Advanced Placement Statistics teacher was not sure about the modeling aspect of this task (Site 3, Tape 8, 1:31:00):

SM: *Where did that equation come from?*
June: *The calculator.*
David: *We did . . . [laughter]*
SM: *What did you do?*
David: *Power regression.*
SM: *What made you do power regression?*
David: *Well, first we plotted the points and we were looking at a power regression with a negative, I don't know, it just fit the . . .*
SM: *What do you think of with power regression?*
David: *Yeah. Well first we tried like exponential. Actually, first I thought maybe it was logarithmic, but the power regression seemed to fit it better.*

The final algebraic models recorded by teachers for their individual investigations at Site 3 are provided in Figure 9.

1) $y = 50*(1/x) + 1.48$ 2) $y = 9.66*x^{(-1/2)}$ 3) $y = 0.073*x^{(-.475)}$ 4) $y = 22.87*x^{(-.49)}$ 5) $y = 11.314/\sqrt{x}$ 6) $y = 15.6x^{(-1/2)}$ 7) $y = 2.695x^{(-.504)}$ 8) $y = 7.43 + -1.5\ln x$

Figure 9. Teachers' mathematical models from Activity 3.4.

All but two of the models (#1 and #8) shared the common general form $y = \frac{\sigma}{\sqrt{x}}$,

where x is the sample size and y is the standard error. The next transcript picks up with the discussion of the interpretation of these models (Site 3, Tape 8, 1:50:30).

SM: *Scott, you just said something about inverse variation.*
Scott: *y varies inversely with the square root of x.*

The facilitator mentioned her observation that teachers frequently were seen “marching through the regressions” on the calculators and suggested that this may not be very reminiscent of mathematical thinking (Site 3, Tape 8, 1:51:00).

SM: *If we're viewing mathematics from a sense-making perspective, and we want our kids to be mathematical sense-makers, then we need to be mathematical sense-makers too, OK? And so, if we have a curve, that looks like that [referring to the graph], does it have a message in it for us?*
Jennifer: *There are things it can't be for sure, like we know it's not linear or quadratic.*
SM: *Yeah, it doesn't look linear, it doesn't look quadratic.*
Beatrice: *It looks exponential to me.*
SM: *It might have looked exponential, because the reasoning for exponential would be what?*

The group discussed conditions for exponential decay and then forayed into inverse variation in the general sense. Then teachers were asked to think about how they might generate a function to model the situation from the tabled data without using linear regression. Collectively, teachers were able to argue for a way to generate the inverse variation model; however, it was fairly evident, and the video confirmed this, that this was not a typical practice for them. In addition to guiding teachers through mathematical modeling “by hand,” the facilitator modeled the use of sliders on the function plot for inverse variation in *Fathom2*. Looking back at the table of data, teachers articulated their understanding of the Central Limit Theorem. Through discussion, the facilitator

attempted to connect what was being learned about sampling to the creation of confidence intervals. Teachers volunteered their thinking for the group and were seemingly able to connect characteristics of normal distributions, standard error, and sampling to make sense of this idea. The conversation was brief but interactive.

Connecting samples with density curves. The professional development concluded with two final activities. First, teachers were asked to match samples taken from populations with given, but different, density curves. The density curves were normal, uniform, skewed, and bimodal. The first matching task involved samples of size 100. Teachers confidently completed the task. The second matching task involved samples of size 10. Teachers were not nearly so confident in these choices and concluded that larger samples contain more information from which to reason.

Generating informed cognitive conflict, connections, and resolution. Finally, the concluding activity involved an exploration of “The Physician’s Health Study” (Zitek, 2006). The investigation was more abbreviated and more teacher-directed at Site 3 than at the other two sites due to time constraints. The general question of interest from the study was, “Do physicians discriminate against patients on the basis of their weight?” More specifically, “Do the doctors who review charts of overweight patients say they would spend the same amount of time with their patients as the doctors who review charts of normal weight patients?” Teachers read the case study and examined the tables of data; they acknowledged that the design of the study seemed reasonable for conclusions to be confidently drawn.

Almost immediately teachers asked for a graphical representation. When asked for their preference, they replied, “histograms.” Side-by-side histograms of the data were projected on the screen for their view. When asked whether they felt that the means of the distributions would be significantly different, about half of the teachers said no, including the Advanced Placement Statistics teacher. They discussed their rationale and another teacher asked to see boxplots. Several teachers mentioned there was overlap among the boxes, suggesting that there would be no significant difference. One teacher asked to move one outlying value in toward the median to see what would happen to the mean. Others asked for numerical comparisons. With a mean difference of 6.7 minutes, many of the teachers felt the difference was small enough and the overlap of the boxplots was sufficient for no significant difference. Figure 10 contains graphical and numerical representations of the data.

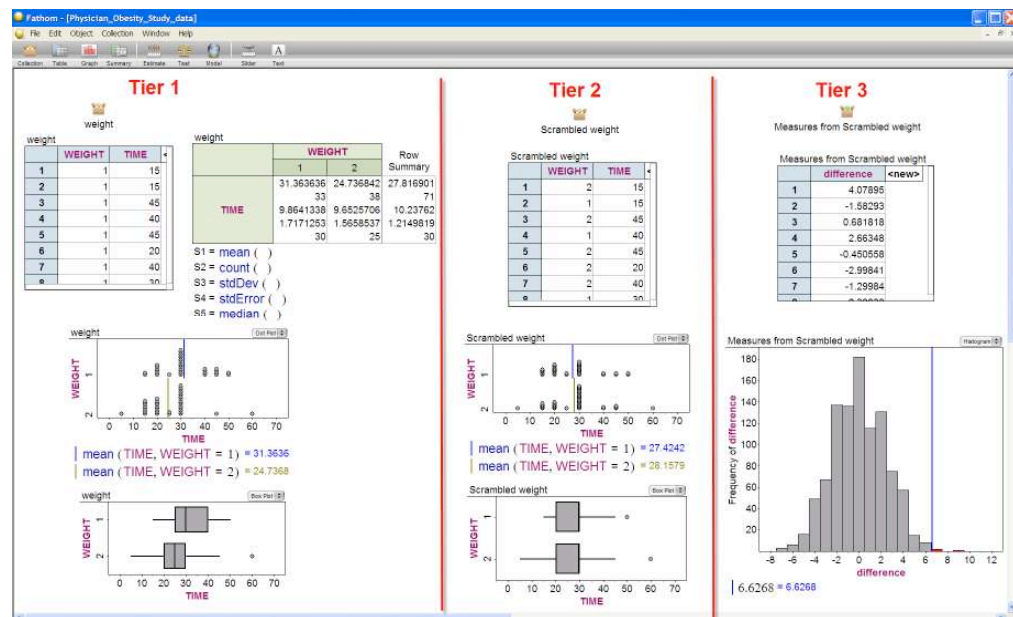


Figure 10. The randomization test for the Physician's Health Study.

The following discussion represents teachers' response to the randomization test result (Site 3, Tape 9, 40:30).

- Blaire: *It happens randomly so little, so if it's just left to chance it doesn't really happen that often.*
- SM: *Less than 5% of the time we see results that unusual. The probability of it happening by chance alone is small; it doesn't happen very often, we can see it. And it did happen. It happened in our experiment. Is it likely due to chance [motioning to the randomization distribution on the screen]?*
- Teachers: *No.*
- SM: *No, less than 5% of the time it is. It's likely not due to chance, it is likely due to something else. In this case, in this context, what is it likely to do with?*
- Scott: *We controlled everything else.*
- SM: *We controlled for everything else.*
- Scott: *It'd have to be those two factors of obesity and normal weight.*
- SM: *Right? Good?*
- Beatrice: *But in all the examples we did though, when we looked at the graphs and there's a bunch of overlap, then it wasn't significant. And when we looked at the graphs, there was the overlap, so that's the reason I said not significant.*
- SM: *Yeah, but see, that's an important thing. When we make these decisions, we're making them on the basis of something. Maybe we're comparing histograms, maybe we're comparing boxplots, maybe we're comparing who knows what, and we're saying, "Yeah, it seems all right to me," or whatever it is we're saying. And the bottom line is that it's not that simple. There are some guidelines, yeah. If the boxes are like this [motioning for very far apart], pretty good indicator there's probably a difference. If the boxes are like this [motioning closer together], a little closer to call. Yeah [responding to question].*
- Alexandra: *I want to make sure I'm correct. If it's a low p-value then it is a significant difference. If it's a high p-value then it's not significant.*
- SM: *Yeah, right because a high p-value, p stands for probability, right? If the probability of seeing a result like that is high, it can happen a lot by chance, then it's probably not due to our treatment. But if it can't happen by chance, it's probably due to something we controlled for.*

It is clear from the transcript that some teachers continued to wrestle with the meaning of the randomization test, its interpretation, and connections to previous representations.

One thing that is not so clear is that a similar thing happened in each of the three sites:

Teachers made predictions on the basis of prior understanding and intuition. For many

teachers, the disequilibrium from this task may have encouraged them to reflect on the ways in which they understood significant difference. At Sites 1 and 2, teachers explored the case in pairs, at site 3, the whole group investigated the question and the facilitator used the software to make various representations as suggested by teachers. At all sites, the highly significant difference in sample means was surprising for many teachers. It seemed to have the effect of causing them to reconsider histogram and boxplot representations as well as the magnitude of differences which might be significant.

The session ended with a debriefing conversation about the final activity and a binder-walk through the materials that teachers would take with them and would have at their disposal for future use. Big ideas for the week were generated and teachers were encouraged to read the articles from the binder they had not yet had time to read, as well as the NCTM Yearbook they had been provided. They were advised that site licenses for *Fathom2* software would be arriving at their schools in the near future and to go and make the world a more statistically-literate place. After a brief break, teachers completed the post-assessment and the session ended.

Looking Back Over the Four-Day Professional Development Programs

By some measure, four days is a brief period during which to influence others' understanding. On the other hand, four days of intensive investigation and support in a technologically-rich environment, as described in this study, may be sufficient to build upon and extend the statistical foundation that practicing teachers enter with, regardless of prior statistical acumen. The evidence presented through the articulation of the enacted learning trajectory in this chapter suggests that conditions can exist or be engineered for

which extending teacher learning of statistics is possible. *Comparing distributions* with specific emphasis on distribution, variability, and sampling distributions and their connections appears to be a viable vehicle, in a technologically-rich environment, for supporting teachers' understanding. Additional evidence will be presented in Chapter V.

Summary of Themes Emerging from Retrospective Analysis of Professional Development Programs

All 60 hours of videotape recorded for this study were reviewed, cataloged, and analyzed by looking for recurring themes, patterns, and characteristics supporting teachers' learning of *comparing distributions*. The following two research questions guided and informed the analysis.

1. How do professional development experiences with resampling techniques and dynamic statistical tools, as described in this study, shape what teachers know about *comparing distributions*?
2. What characteristics of professional development for high school mathematics teachers contribute to their understanding of *comparing distributions*?

Themes or characteristics that were ongoing and evident at each of the professional development sites and those that appeared influential for productive teacher learning of *comparing distributions* are included in Table 17. Table 17 is an attempt to coordinate the complexities associated with statistical professional development for high school mathematics teachers as described in this study. Evidence for each of the characteristics in the table can be found in the transcribed or described episodes earlier in the chapter. The evidence presented in this chapter is but a small sample of that analyzed.

As seen in Table 17, making statistical arguments, privileging sense-making, and continually navigating within and among multiple-representations, appear to be common factors representing characteristics of professional development as well as experiences with resampling techniques and dynamic statistical tools associated with supporting teachers' understanding of *comparing distributions*. The use of accessible and engaging investigations, the establishment of an intellectually safe environment, and intentionally sharing authority for the knowledge in the classroom were all characteristics of professional development which appeared important for supporting teachers' understanding of *comparing distributions*. Access to multiple dynamic statistical software tools in conjunction with investigations designed to utilize resampling techniques, most notably, the randomization test, appear influential in helping to shape what teachers come to understand about *comparing distributions*. It seems unlikely that some of the revelations and connections made by teachers during the professional development experience would have been possible in the absence of this technology utilized in the context of resampling to make comparisons.

Coordinating Content, Technology, and Professional Development Intervention Characteristics

While attempting to capture the many intricacies of 60 hours of content-focused, technology-rich professional development, a large challenge was how and what to present in order to both provide a flavor for the nature of the professional development and simultaneously a rigorous, thoughtful, thorough analysis of the enacted learning trajectory. Much like a statistical analysis, the search is for the signal amid the noise. The analytical process required multiple passes through the data, each time trying to further

Table 17

Summary of Factors Potentially Impacting Teachers' Understanding of Comparing Distributions

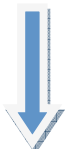
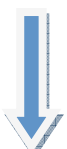
Characteristics of Professional Development							
			Use of Dynamic Technology & Resampling				
<i>Accessible & Engaging Investigations</i>	<i>Intellectually Safe Environment</i>	<i>Sharing Authority for Knowledge</i>	<i>Making Statistical Arguments</i>	<i>Sense-making</i>	<i>Multiple Representations</i>	<i>Dynamic Statistical Software</i>	<i>Resampling</i>
Interesting contexts	Non-evaluative stance	Not just searching for "right answers"			Fathom2		
Building on prior knowledge	Responsibility for sharing ideas and thinking	Obligation to make thinking visible and public	Reasoning was privileged		CMP-Tools		
Connections to previous and future activities	Support for ideas and conjectures balanced with need for evidence	Learning together	Divergent thinking encouraged	Making predictions	Moving between graphical, tabular, symbolic representations	Randomization Distribution	
Potential use with high school students	Careful listening	Appreciating different points of view and experience	Evidence for statements required	Coordinating representations and perspectives	Histogram, dotplots, boxplots, function plots, summary tables	Sampling Distributions (of the mean)	
Connections to current practice--extensions	Equitable airtime		Connections to mathematical reasoning	Cognitive dissonance	Contrasting shapes, centers, variability	Visibility of process	
	Humor		Articulating thinking verbally and in writing			Scaffolding with technology	
	Vulnerability						

- = Characteristics of professional development
 = Use of dynamic technology and resampling
 = Overlap across both characteristics of professional development and use of dynamic technology and resampling

distill and refine the various signals. Table 18 provides a global view of the evolution of the statistical content, use of technology, and characteristics of professional development that emerged from the retrospective analysis of the video-tape of the professional development program in this study. This representation is an attempt to synthesize and extract important messages from a large and complex data source. A great deal of detail

Table 18

Summary of the Enacted Evolution of Content, Technology Use, and Characteristics of Professional Development

Day	Big Ideas	Evolution of Big Ideas	Evolution of Technology Use	Ongoing Characteristics of PD
1	- Experimental Design - Characteristics of Distributions - Measures of variability - Introduction to navigating in <i>Fathom2</i> - Properties of Good Estimators - Making Statistical Arguments	Experimental Design, Characteristics of Distributions & Properties of Estimators  Randomization Testing & Establishing Cause and Effect  Building Sampling Distributions & Making Connections	- Low-tech to begin - Modeling navigation in <i>Fathom2</i> & <i>CPMP-Tools</i> during investigations - Stat & List features of TI-84 Plus	- Establishing a safe culture - Providing provocative contexts for investigation - Encouraging statistical arguments - Privileging sense-making - Challenging & supporting teachers' existing conceptions - Discouraging disposition toward "right answers" - Accessible tools to support investigation and conjecturing - Encouraging multiple representations - Using dynamically-linked representations - Big ideas connecting activities & trajectory
2	- Randomization Distribution - Introduction to <i>CPMP-Tools</i> - Introduction to <i>Fathom2</i> - Samples and sampling - Bias - Building the randomization distribution mechanism in <i>Fathom2</i> - Cause & Effect		<i>CPMP-Tools</i> for randomization testing - Tours of <i>Fathom2</i> - More modeling of <i>Fathom2</i> during investigations	
3	- Cause & Effect (cont.) - Using the randomization test to compare difference in means and medians - Properties of Normal Distribution - Making connections		Building and using the randomization distribution mechanism in <i>Fathom2</i>	
4	- Function plotting and the Normal Density Function in <i>Fathom2</i> - Sampling, sampling distributions and modeling the Central Limit Theorem - Matching samples to density curves - Checking intuition and skills		- Function plotting in <i>Fathom2</i> - Building and using sampling distributions in <i>Fathom2</i>	

is lost; however, the loss in detail is replaced by the potential future usefulness of the design characteristics inferred in the representation.

The description, evidence, and synthesis from the enacted learning trajectory will be referred to again in Chapter VI. Chapter V will present the results from teachers' pre- and post- written assessments, pre- and post-interviews, and teachers' written reflections from the professional development program.

CHAPTER V

RESULTS

The chapter is organized into the following major sections: Content Assessment Results, Interview Results, and Written Reflection Results. Each section has its own summary and a brief segue to Chapter VI is provided at the end of the chapter.

Content Assessment Results

A written pre-assessment was administered to teachers approximately two months prior to the professional development described in this study. A written post-assessment was administered to the same group of teachers on the final day of the professional development program at each of the three professional development sites. The pre- and post-assessments consisted of 21 statistical questions or tasks, a combination of multiple choice, multiple choice with explanation required, and open-ended items. The tasks were grouped into 10 item clusters (referred to as items), each representing one of three “big ideas” of *comparing distributions*, around which this study is centered (Table 19). The assessment framework was presented in Chapter IV (Figure 5). Pre- and post-assessment instruments may be found in Appendices A and B. The theoretical framework from Chapter III was used to create item-specific four-level rubrics (see Appendix F) for scoring the assessment items.

Of the 10 items on the pre- to post-assessments, eight may be considered parallel items that can be legitimately compared. Table 19 displays the correspondence between the 21 assessment tasks and the 10 item clusters.

Table 19

Correspondence Between Assessment Tasks and Item Clusters

Question or Task (T)	Item Cluster (I)	Pre/Post Parallel?
1	1	Yes
2	2	Yes
3	3	Yes
4a, 4b, 5, 7	4	Yes
6	5	Yes
8	6	No
9	7	Yes
10, 11, 12	8	Yes
13	9	No
14, 15, 16, 17, 18, 19, 20	10	Yes

Two tasks, T8 and T13, should not be considered parallel. Even though both pre- and post-assessment T8 and T13 were each included to assess teachers' understanding of simulation and experimental design, respectively, differences in the tasks could be considered significant enough as to make direct comparisons problematic. The results of the tasks convey important information for the purpose of assessing teachers' understanding of *comparing distributions*, but the interpretation of statistical change from pre- to post-assessment by item is done with caution.

For completeness, the aggregate results from the pre- and post-assessments are reported in two ways. First, aggregate results are reported for the complete set of ten items on the assessments. Additionally, aggregate results containing only the eight parallel items are presented. Individual items were scored using four-level rubrics. Aggregate scores are in the range of 0–4, representing the average level of response across items. Interpretation of the aggregate score is associated with the framework in

Table 5 presented in Chapter III. Additional item comparisons are made for all 56 persons completing the pre- and post-assessments and change-scores from pre to post.

Aggregate scores were similar with or without the two non-parallel items included in the analysis. The mean pre- and post-scores for the eight parallel items were slightly lower than the aggregate including all ten items; however, gain-scores were not significantly different from each other. The mean gains from pre- to post-assessment were approximately 0.89 points (effect size 1.83) for the 10-item comparison and 0.86 points (effect size 1.58) for the eight parallel-item comparison. This suggests that the gain seen across all ten items was consistent with the gain in parallel items.

Establishing Comparability Across Sites

Data from the pre-assessment for the 56 teachers were analyzed by site in which the professional development occurred in order to determine whether the three groups of teachers were initially comparable. As suggested in Table 20 and Figure 11, the differences between the mean scores for each group were not significant. Because the assumptions for one-way ANOVA were questionable in this case, a Kruskal-Wallis test was run with $k = 3$, $n_1 = 31$, $n_2 = 9$, $n_3 = 16$. Mean ranks were Site 1—32.2, Site 2—20.4, Site 3—25.9; $H = 4.24$, $df = 2$, $p = 0.12$. Therefore, there was not strong evidence to conclude that the pre-assessment mean scores of the three groups were unequal. Item-by-item comparisons are summarized by site in a later section of this chapter for comparison purposes.

Table 20

Aggregate Pre-Assessment Scores Across Sites

	Site 1	Site 2	Site 3	Overall
<i>M</i>	1.97	1.64	1.89	1.89
<i>Mdn</i>	1.95	1.65	1.78	1.85
<i>SD</i>	0.46	0.38	0.55	0.48
<i>n</i>	31	9	16	56

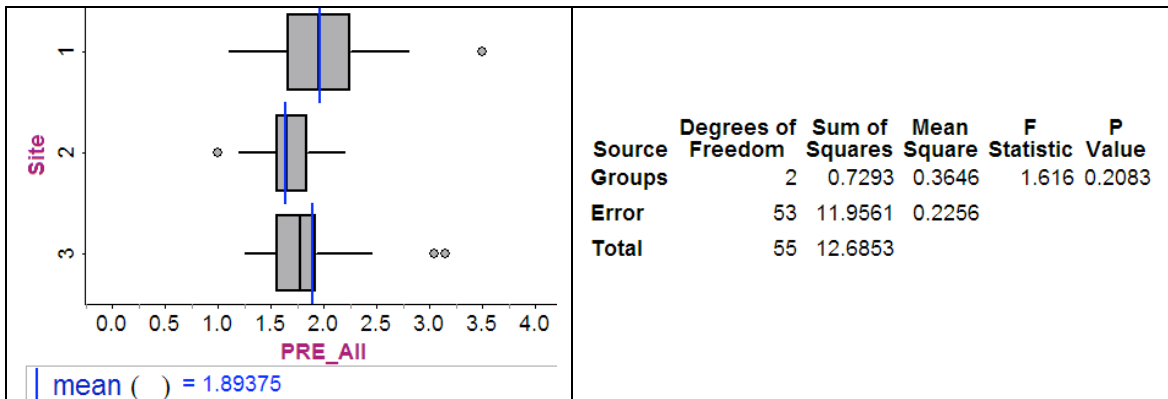


Figure 11. Comparison of aggregate pre-assessment scores across sites with ANOVA results.

Aggregate Results (All 10 Items)

Pre-assessment results are presented, followed by post-assessment results and then gain-score results. All of the sections will report results for the complete set of 10 item clusters on the assessment. Following the overall gain-score results, comparisons using just the eight parallel item clusters are presented. A summary is provided at the end of the section.

Pre-Assessment Results

Table 20, presented previously, displays the by-site and overall results for the pre-assessment when scores for all 10 item clusters were averaged for each individual teacher

in the project. Though there is some variation in pre-assessment scores across sites, all sites recorded initial mean scores of less than 2 on a four-level scale. Using a one-way ANOVA and Kruskal-Wallis test to test for differences in average scores, no significant differences across regions were identified (see Figure 11). The pre-assessment scores shown in Table 21 were further disaggregated by item in order to determine the relative difficulty of each item for this group of teachers.

Table 21

Distribution of Pre-Assessment Scores by Item

Item	Task(s)	0	1	2	3	4	<i>M</i>	<i>Mdn</i>	<i>SD</i>	Relative Difficulty ^b
1	1	0	4	38	12	2	2.21	2	0.624	8
2	2	1	27	13	11	4	1.82	1.5	1.011	5
3	3	0	31	0	0	25	2.34	1	1.505	9
4	4,5,7	11	21	14	6	4	1.48	1	1.144	2
5	6 ^a	4	9	22	14	7	2.47	2.5	1.020	10
6	8	4	17	23	11	1	1.79	2	0.909	4
7	9	1	9	35	9	2	2.04	2	0.738	7
8	10,11,12	5	10	32	6	3	1.86	2	0.923	6
9	13	6	12	33	5	0	1.66	2	0.793	3
10	14-20	10	32	7	3	4	1.27	1	1.053	1

^a For this table, scores from task 6 were determined using the greatest integer function. See the discussion on task 6 in the current chapter for additional details on the scoring of this item.

^bRelative difficulty was determined by the mean score and further informed by the median score. The most difficult item was determined by the lowest mean score and was rated 1; the least difficult item was rated 10.

None of the items on the pre-assessment produced mean scores higher than 2.47. On six of 10 items, most teachers scored at Level 2, while on the remaining four items, most teachers scored at Level 1. The relative difficulty index suggests that Item 10, requiring understanding sampling distributions and the Central Limit Theorem was the most challenging for teachers, while Item 6, requiring teachers to reason from boxplot representations to compare distributions was the easiest. More details about individual tasks are presented with the item-by-item analysis in a later section of this chapter.

Post-Assessment Results

Table 22 displays the by-site and overall results for the post-assessment when scores for all 10 item clusters were averaged for each individual teacher in the project. Though there is some variation in post-assessment scores across sites, all sites recorded mean scores greater than or equal to 2.56 on a four-level scale, with a mean post-assessment score of 2.78 for the group. Using a one-way ANOVA to test for differences in average scores, no significant differences across regions were identified (see Figure 12). Because the assumptions for one-way ANOVA were questionable in this case, a Kruskal-Wallis test was run with $k = 3$, $n_1 = 31$, $n_2 = 9$, $n_3 = 16$. Mean ranks were Site 1—28.9, Site 2—22.0, Site 3—31.4; $H = 1.96$, $df = 2$, $p = 0.3753$. Therefore, there was not strong evidence to conclude that the post-assessment mean scores of the three groups were unequal. The post-assessment scores shown in Table 23 were further disaggregated by item in order to again determine the relative difficulty of each item for this group of teachers.

Table 22

Aggregate Post-Assessment Scores Across Sites (10 Items)

	Site 1	Site 2	Site 3	Overall
<i>M</i>	2.81	2.56	2.84	2.78
<i>Mdn</i>	2.85	2.70	3.03	2.88
<i>SD</i>	0.473	0.593	0.622	0.54
<i>n</i>	31	9	16	56

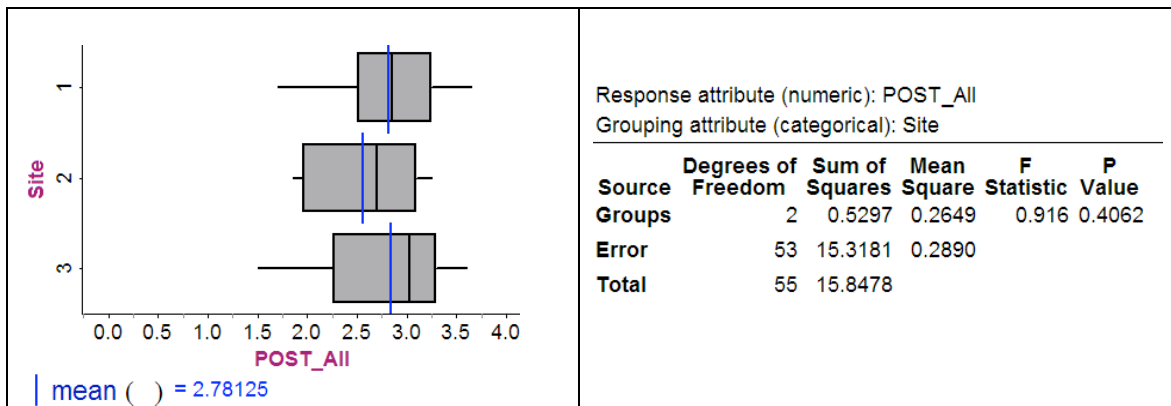


Figure 12. Comparison of aggregate post-assessment scores across sites with ANOVA results.

Table 23

Distribution of Post-Assessment Scores by Item

Item	Task(s)	0	1	2	3	4	<i>M</i>	<i>Mdn</i>	<i>SD</i>	Relative Difficulty ^b
1	1	0	0	25	23	8	2.70	3	0.711	6
2	2	0	12	13	13	18⁺⁺⁺	2.66	3	1.149	5
3	3	0	13	0	0	43⁺⁺⁺	3.30	4	1.28	10
4	4,5,7	0	4	7	20	25⁺⁺⁺	3.18	3	0.917	9
5	6 ^a	0	1	16	35⁺	4	3.03	3	0.599	7
6	8	0	5	6	21	24⁺⁺	3.14	3	0.943	8
7	9	0	1	22	29⁺	4	2.64	3	0.645	4
8	10,11,12	0	3	26	20	7	2.55	2	0.784	3
9	13	2	2	33	14	5	2.32	2	0.834	2
10	14-20	2	15	17⁺	9	13	2.29	2	1.201	1

^a For this table, scores from task 6 were determined using the greatest integer function. See the discussion on task 6 in the current chapter for additional details on the scoring of this item.

^b Relative difficulty was determined by the mean score and further informed by the median score. The most difficult item was determined by the lowest mean score and was rated 1; the least difficult item was rated 10.

Each ⁺ represents a one-level modal shift from the pre-assessment

None of the items on the post-assessment produced mean scores lower than 2.32.

Mean scores increased for all 10 items on the post-assessment compared to the pre-assessment and median scores increased for eight of 10 items. On four of 10 items, most teachers scored at Level 4, on two items most teachers scored at Level 3 and on the

remaining four items most teachers scored at Level 2. Positive modal shifts occurred on seven of 10 items, with three items showing modal shift of +3 levels. The relative difficulty index suggests that Item 10, requiring understanding sampling distributions and the Central Limit Theorem remained the most challenging for teachers, while Item 3, requiring teachers to reason about variability while considering sample size was the easiest. Items 2 and 10 retained their relative difficulty indices from the pre-assessment whereas all other items shifted in terms of relative difficulty.

Gain-Score Results

Table 24 displays the by-site and overall results for the pre- to post-assessment gain-scores. Though there is some variation in gain-scores across sites, all sites recorded gain scores greater than or equal to 0.85 on a four-level scale, with a mean gain-score of 0.89 for the entire group. Using a one-way ANOVA to test for differences in mean gain-scores, no significant differences across regions were identified (see Figure 13). Because the assumptions for one-way ANOVA were questionable in this case, a Kruskal-Wallis test was run with $k = 3$, $n_1 = 31$, $n_2 = 9$, $n_3 = 16$. Mean ranks were Site 1—27.3, Site 2—29.1, Site 3—30.4; $H = 0.39$, $df = 2$, $p = 0.8228$. Therefore, there was not strong evidence to conclude that the mean gain-scores of the three groups were unequal.

Table 24

Pre- to Post-Assessment Gain-Scores (10 Items)

	Site 1	Site 2	Site 3	Overall
<i>M</i>	0.85	0.92	0.95	0.89
<i>Mdn</i>	0.85	1.00	0.88	0.88
<i>SD</i>	0.38	0.68	0.56	0.49
<i>n</i>	31	9	16	56

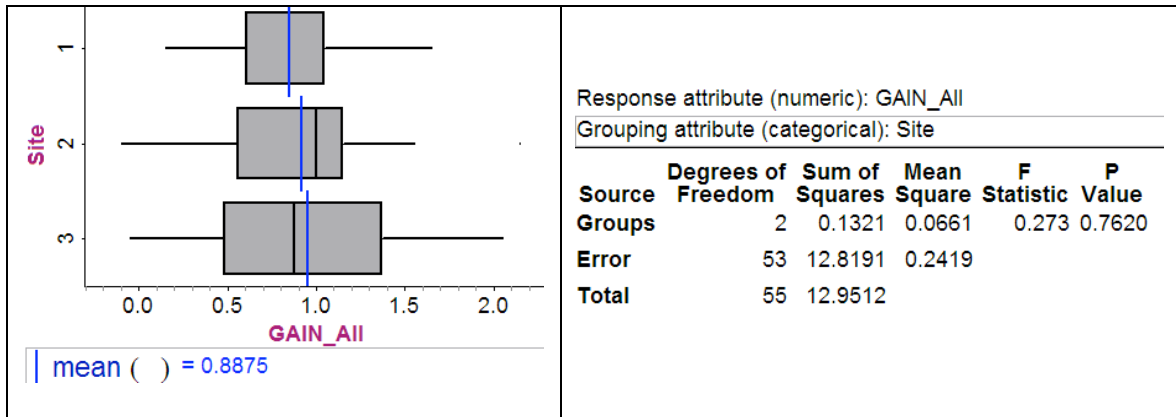


Figure 13. Comparison of aggregate gain-scores across sites with ANOVA results.

Figure 14 provides visual and numerical evidence suggesting a positive shift in the distribution of scores from pre- to post-assessment. Mean gain-scores of 0.89 are significantly greater than 0 ($p < 0.0001$).

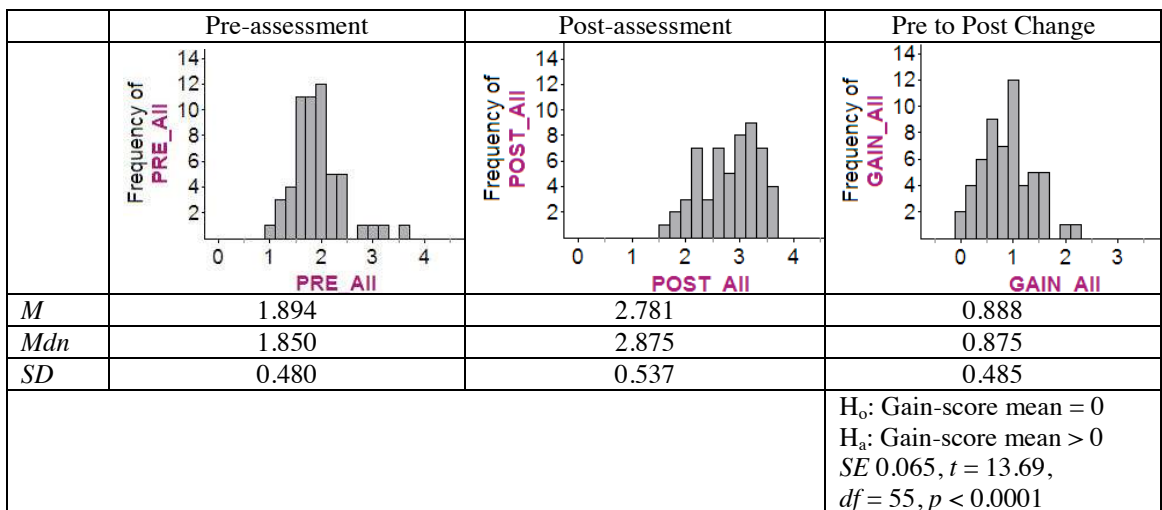


Figure 14. Comparison of pre-, post-, and gain-score distributions and matched pairs t -test results (10 items).

Aggregate Results (Eight Parallel Items)

A number of comparisons are provided to illustrate the similarities and differences between the results of the content assessments when all 10 items are

considered versus when only the eight parallel items are included. Figure 15 provides visual and numerical evidence suggesting a positive shift in the distribution of scores from pre- to post-assessment. Mean gains-scores of 0.86 are significantly greater than 0 ($p < 0.0001$). The 0.86 mean of the gain-score distribution is slightly lower than the 0.89 mean of the 10-item distribution of gain-scores, whereas the standard deviation of 0.54 is slightly greater than the 0.49 of the 10-item distribution.

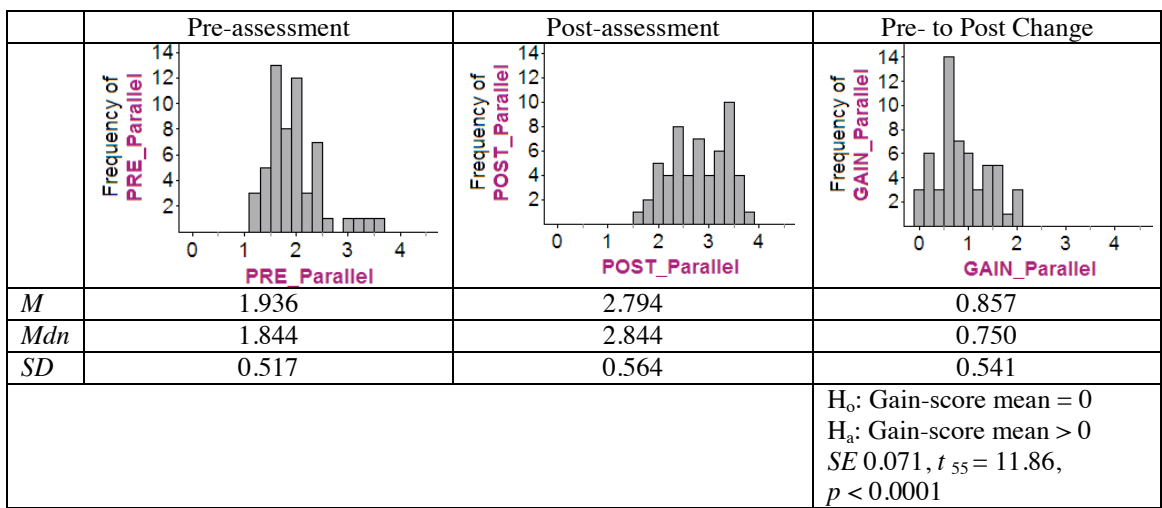


Figure 15. Comparison of pre-, post-, and gain-score distributions and matched pairs t -test results (eight parallel items).

The correlation between parallel gain-scores and total gain-scores is quite high ($r = 0.95$) with a strong linear pattern (Figure 16). The corresponding residual plot supports the linear model. This suggests that the information provided by both the complete 10-item analysis is consistent with that from the eight parallel-item analysis.

When the post-assessment item means are plotted against the pre-assessment item means for all 10 items as in Figure 17, with the line $y = x$, it becomes clear that the performance on the all items improved on the post-assessment. Furthermore, the relatively low post-assessment mean on Q13 (Item 9) is contrasted with the relatively

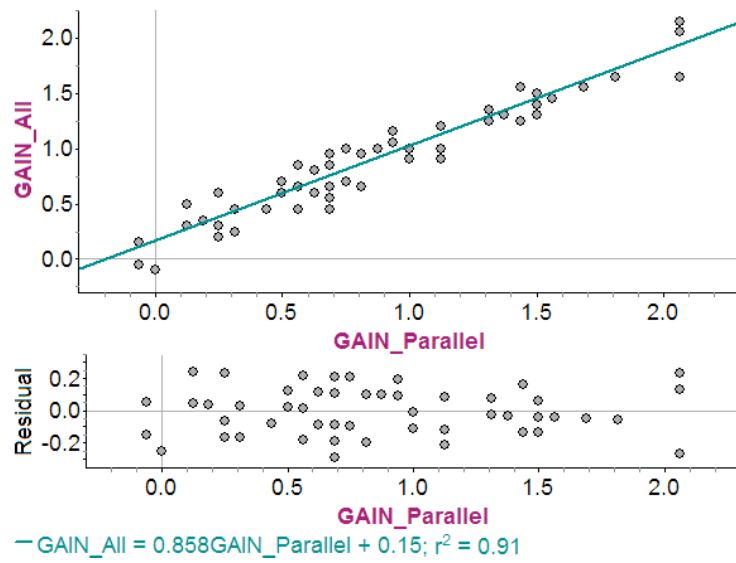


Figure 16. Gain-scores for 10 items compared to gain-scores for 8 parallel items.

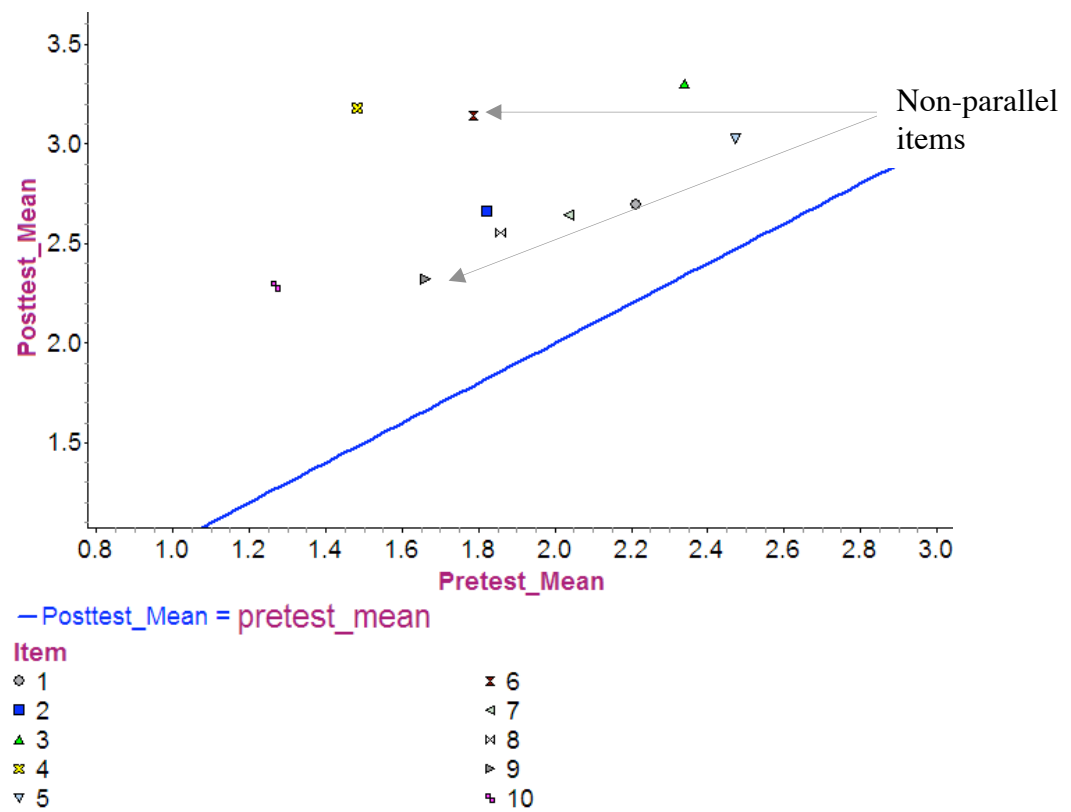


Figure 17. Item-by-item post-assessment means against pre-assessment means.

high post-assessment mean on Q8 (Item 6). The results visually and numerically balance each other which is why the previous analyses with 8 parallel items and all 10 items do not appear appreciably different from one-another.

The results presented thus far support the position that for purposes of comparison, using all 10 items for analyses is defensible. Arguably, the analyses could proceed with or without the non-parallel items. For purposes of the next comparisons, all 10 items are included.

Aggregate Results: Additional Comparisons

Figure 18 displays the number of statistics courses taken by teachers, and pre-, post-, and gain-score distributions disaggregated by number of statistics courses taken. Most teachers (46 of 56) had taken one or more statistics courses; however, differences among mean scores on the pre-assessment, as a function of previous statistical coursework, were not statistically significant. Previous coursework in statistics appeared to make little difference in teachers' responses to the written assessments used in this study. Statistic course-taking apparently had minimal influence on teachers' abilities to respond to the questions on the assessments. This may indicate that previous coursework in statistics failed to leave much accessible residue in order for teachers to reason statistically or possibly that the items on the assessment were markedly different from their prior statistical coursework or experience. The positive shift in distributions for teachers, regardless of prior statistical training, furthermore suggests that the professional development experience during this research study may be broadly applicable across the spectrum of heterogeneously statistically-knowledgeable teachers.

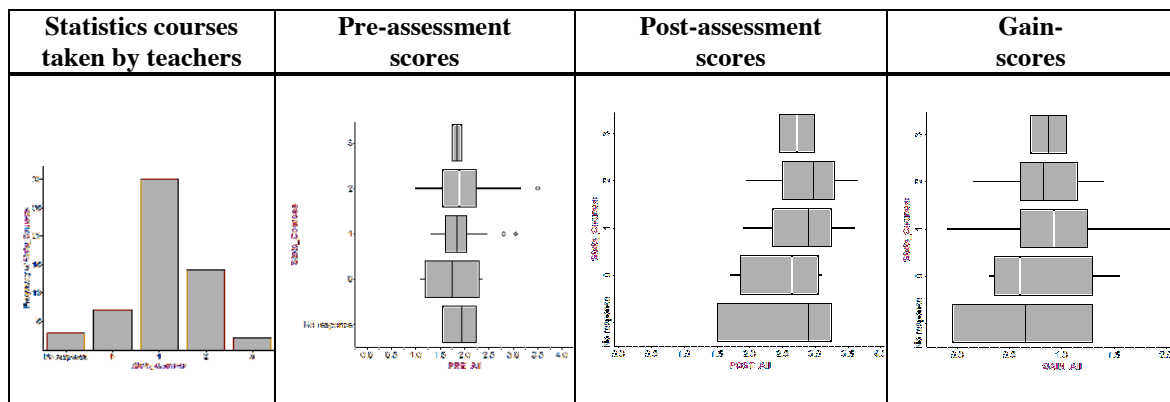


Figure 18. Comparisons of pre-, post-, and gain-score distributions by number of statistics courses taken.

Table 25 presents pre-, post-, and gain-scores for each assessment item by professional development site. This table is provided for completeness, but also to showcase the remarkable similarities across the performance of the three groups of teachers in this study on the assessments. Gain-scores for Item 2 and Item 8 differed significantly across regions as did post-assessment scores for Item 1. Upon further investigation, a clear pattern with explanatory power could not be determined. For example, with respect to Items 2 and 8, both dealing with understanding variability, Item 2 gain-scores were significantly higher for Site 2, whereas Item 8 gain-scores were significantly lower for Site 2.

The item-based differences across sites disappear as items are aggregated into big idea clusters of distribution, variability, and sampling distributions. When mean pre-, post-, and gain-scores were analyzed by big ideas of distribution, variability, and sampling distribution to determine whether differences between sites could be detected, using one-way ANOVA with $\alpha = 0.05$, none of the comparisons were found to be significant. This evidence suggests that with respect to the three big ideas investigated in this study, teachers across three sites entered with comparable prior knowledge, left with

Table 25

Content Assessment Mean Pre-, Post-, and Gain-Scores by Professional Development Site

Task	Site	Pre-assessment	Post-assessment	Gain
1 Item 1	1	2.23	2.74*	0.52
	2	2.00	2.11*	0.11
	3	2.31	2.94*	0.63
2 Item 2	1	2.06	2.52	0.45*
	2	1.22	3.00	1.78*
	3	1.69	2.75	1.06*
3 Item 3	1	2.35	3.42	1.06
	2	2.33	3.33	1.00
	3	2.31	3.06	0.75
4a, 4b, 5, 7 Item 4	1	1.42	3.26	1.84
	2	1.22	2.89	1.67
	3	1.75	3.19	1.44
6 Item 5	1	2.55	3.00	0.45
	2	2.11	2.94	0.83
	3	2.53	3.13	0.59
8 Item 6	1	1.84	3.32	1.48
	2	1.33	2.66	1.33
	3	1.94	3.06	1.13
9 Item 7	1	2.16	2.68	0.52
	2	1.67	2.44	0.78
	3	2.00	2.69	0.69
10, 11, 12 Item 8	1	1.97	2.52	0.55*
	2	1.89	2.11	0.22*
	3	1.63	2.88	1.25*
13 Item 9	1	1.81	2.32	0.52
	2	1.56	2.11	0.56
	3	1.44	2.44	1.00
14-20 Item 10	1	1.29	2.35	1.06
	2	1.11	2.00	0.89
	3	1.31	2.31	1.00
Distribution	1	2.09	2.92	0.83
	2	1.75	2.60	0.85
	3	2.15	2.98	0.84
Variability	1	2.13	2.82	0.69
	2	1.81	2.81	1.00
	3	1.88	2.90	1.02
Sampling Distribution	1	1.65	2.67	1.02
	2	1.33	2.26	0.93
	3	1.56	2.60	1.04

*One-way ANOVA test of equality of means by site were significant at $\alpha=.05$.

comparable knowledge, and gained comparable knowledge from pre- to post-assessment. A big idea summary table is provided following the item-by-item analyses which are presented next.

Item-Specific Analyses

Analyses of each item (see Table 19 for correspondence between assessment tasks or questions and item clusters) that contributed to a teacher's overall score on the pre-and post-assessment are presented here. For each item, the distribution of the pre-assessment scores, post-assessment scores, and gain scores are included. Matched pairs gain-scores are tested to determine whether they are significantly greater than 0 with $\alpha = 0.001$ to compensate for the numerous comparisons in this study. By reducing the alpha-level, the likelihood of making a Type I error is reduced. Wilcoxon matched pairs signed ranks tests and randomization tests were evaluated for each comparison. Results of both tests were consistent, and p -values from the randomization tests are presented in tables that follow. The number of teachers whose scores decreased, remained the same, or increased is documented. Each item analysis concludes with a discussion and interpretation of the per-item comparisons. In some cases a more detailed discussion is included due to the nature of the item or analysis required. When individual teacher's responses are reported, a code of Teacher or T plus a two- to four-digit number is used to represent the teacher (e.g., Teacher 71 or T71 refer to teacher number 71). In the narrative, language will include "Teacher 71," whereas when referring to a response on a pre- or post-assessment, T71 will be used. Furthermore, the code (Pre, T71) refers to Teacher 71's written pre-assessment response. Similarly, (Post, T71) refers to Teacher 71's written post-assessment response.

A scoring rubric was used for each of the items; the scoring rubrics may be found in Appendix F. Most of the scoring involved the interpretation and coordination of open-ended written responses. Scoring of this type can be wrought with challenges. Even when following well-defined rubrics, some responses were difficult to categorize. For this study, the following decision rule was implemented: in the event that categorization was not clearly defined, the score was chosen as the lesser of two choices. The rationale for the decision was for the results to err on the side of conservative estimates rather than artificially-inflated estimates. The same decision rule was used for the pre-assessment and post-assessment, thus the scoring mechanism was consistent. The item analyses presented next are organized into sections; each section represents one of the big ideas presented in Figure 5 in Chapter IV.

Analysis of Item 1

Item 1 corresponds to Task 1 on the pre- and post-assessments. Item1 presented two side-by-side histograms of average life expectancies for women (pre-assessment) or men (post-assessment) from regions in Africa and Europe. Teachers were asked to describe similarities and differences between the two distributions and to conclude something about life expectancies for women (men) in the two countries. The life expectancies were said to have been computed for various regions within each country. Figure 19 displays the distribution of pre-assessment, post-assessment, and gain-scores for Item 1. It also includes a cross-tabulated display and summary of the ways in which teachers' scores changed from pre to post on this item.

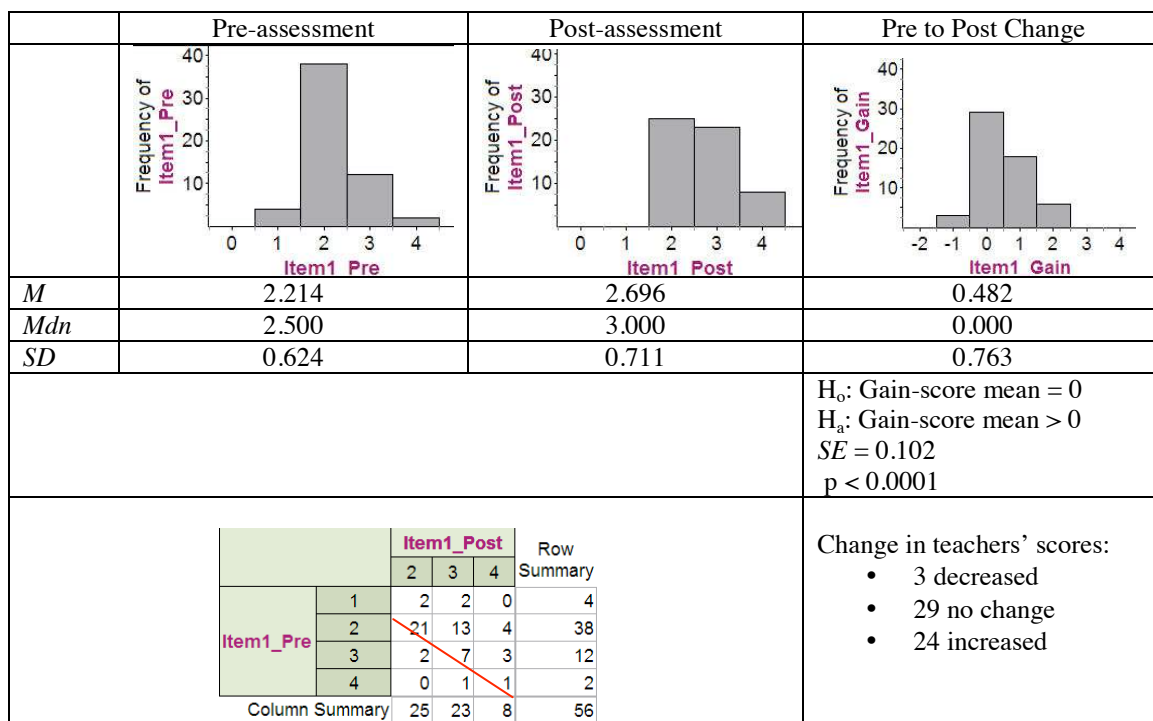


Figure 19. Item 1: Pre-, post-, and gain-score comparisons.

At the time of the pre-assessment, it was hypothesized that this first task would and should be accessible to all teachers and a good starting place for an assessment. Scores on the pre-assessment confirmed that with a mean of 2.21, this item was the third easiest for teachers. An examination of the pre- and post-score distributions reveals a positive shift from pre- to post. The shift is further evidenced by the matched pairs gain-score significantly greater than 0.

Because of the context of the item and teachers' prior knowledge of reading graphs, this item is illustrative of the idea that in a sense-making environment, nearly all teachers could achieve a basic level of success on this task. It was clear from their responses that teachers knew generally how to read the graphs and interpret the information prior to the professional development program. What they chose to write about and whether they did so from a statistical perspective was another matter. This item

was one of several that were particularly challenging to score. The examples in Table 26 may illuminate this point. There was great variability among responses and assessing level of understanding on this task was particularly challenging. Several cases of teacher responses are useful to illustrate the comparison from pre- to post-assessment on this item.

Table 26

Item 1: Sample Responses and Scores

Code #	Pre-assessment	Score	Post-assessment	Score
79	<i>They both show a definite peak. African women tend to "pass" at a younger age although there are apparently some factors that can effect this greatly. The majority of women would have a longer life if they lived in Europe, and although the mean is lower in Africa, there are women who do "make it" to a ripe old age in Africa.</i>	1	<i>1) One mode; 2) Both are skewed although in opposite directions. African men, on average, live shorter lives; life span in Africa has a larger range than that in Europe. Approx 1/3 of both die between 55 and 70.</i>	3
02	<i>They are both skew to one end. The mean life expectancy is much higher for European women. In general, European women have a greater life expectancy then Africa</i>	2	<i>Both have skewed data (draws smooth curves above distributions). European men have a higher mean & median and a smaller range, their data is skewed left whereas the African men have a lower mean & median because the data is skewed right. Men in Europe have a higher life expectancy on average than men in African, although the variance of life expectancies in African men is greater.</i>	4
21	<i>They are both unimodal. The center of the Europe distribution is much higher. The spread of the Africa distribution is much greater. The life expectancy of an African woman is quite variable, but tends to center around 50. The life expectancy of a European woman is much less variable, and centers around the late 70's.</i>	4	<i>They are both unimodal and slightly skewed. The distribution for African men is lower than that for European men. European men have less variability and the distribution is skewed left while African men are skewed right. In general, European men live longer, but this may not be a statistically significant difference. The variability of African men's life expectancy is higher.</i>	4

In the case of Teacher 79, a shift from reasoning from largely contextual cues to attending to more statistical aspects of the distribution such as center, spread, and shape can be seen. Teacher 02 provides a pre-assessment example of attending to shape and center, but with no mention of variability. At post-assessment, the teacher more clearly uses statistical language for making comparisons, addresses center, spread, and variability to some extent and refers to the a higher life expectancy “on average,” suggesting an appreciation of variability. Teacher 21 scored 4 on both assessments, but even so, at post-assessment the language of “European men live longer, but this may not be a statistically significant difference,” may be evidence of the impact of Activity 4.2: The Physician’s Health Study on this teachers’ understanding. This is clearly speculative, but given that Teacher 21 was one of the teachers interviewed for this study and also one of the teachers during Activity 4.2 who voted for “no significant difference” between groups and later learned that the difference was highly significantly different (very low p -value), it may be the case that Activity 4.2 caused this teacher to be conservative in his statement.

As a group, positive change from pre- to post-assessment was seen on this item which included more attention to issues of center (mean, median), spread (range, standard deviation, variance), shape (skewed, normal, bimodal, unimodal), and the use of appropriate statistical language. Only nine teachers referred to the shape of the distributions as skewed on the pre-assessment, whereas on the post-assessment 25 teachers used the language of skewed left or right to refer to the shape of the distributions. Additional change in understanding of statistical language and concepts is further discussed in the analysis of the Item 4 later in this chapter. Teachers appeared

more proficient in their ability to appropriately estimate means and medians of distributions represented as histograms both graphically and numerically and interpret those measures in reasonable ways.

Several factors may have kept teachers from scoring higher on this item. Firstly, as for all of the items, teachers did not have access to the scoring rubric, so they were attending to things they deemed important rather than what the researcher was hoping to see. Secondly, the distributions being compared were mean life expectancies for regions in Europe and Africa. Though this was stated in the task, none of the teachers attended to the fact that the mean of the means might involve a weighted mean comparison in the event that all sample sizes were not equal. It was not the intent of the task to necessarily attend to that aspect; however, it was hypothesized that at least some teachers might address the issue when concluding something about the distributions. For scoring purposes, regions were viewed as equally weighted, even though that is not necessarily the case. Thirdly, never during the professional development program were teachers asked to specifically respond to a task like Item 1 in writing or otherwise. Routinely distributions were being compared in various ways, but it was not the case that specific critiques or opinions were privileged over others or that teachers were directed toward an answer to compare and contrast questions in specific ways. Finally, to score a 4 on this item, a teacher needed to attend to shape, center, and spread in some numeric or graphical way; s/he needed to use statistical language appropriately, and state a conclusion that was correct and not deterministically stated. Meeting the final criterion was often problematic for many participants. An example of a partial conclusion that scored a 4 is “European men live longer on the average. Health care is probably better in Europe” (Post, T2371).

The language of “on average” suggests a stochastic perspective and the contextual statement suggests an intentional connection of the comparison of the histograms to the real situation from which the task originated. Many teachers were “almost there.” Arguably the rubric for this item is picky. On the other hand, to measure change that is desirable, standards must be sensitive to the desired outcome.

Given that no teacher scored less than 2 on this item on the post-assessment and that gains-scores were significantly greater than 0 for the group even when this item was relatively easy for teachers at the time of the pre-assessment, a change from 2.21 to 2.70 (effect size 0.63) still represents an improvement in teachers’ understanding of distributions when comparing histograms.

Analysis of Item 2

Item 2 corresponds to Task 2 on the pre- and post-assessments. Item 2 asked teachers to explain how they would decide which of two events is more unusual. The context was summer (winter on post-assessment) temperatures in cities X and Y. The events were a 90 (5) degree summer day in city X or a 90 (5) degree winter day in city Y. The post-assessment item was analogous to the pre-assessment. A strong response to this item would require attention to the variation in temperatures in each city in order to discern from the distribution of temperatures, the likelihood of a temperature such as 90 (5) degrees; it would be insufficient to reason only from the average temperature. When used by Makar (2004) in her dissertation study, this item resulted in 17% of her students recognizing the need to attend to issues of variability at the pre-test and 69% at the post-test. In the present study, responses scored as 1 were those attending only to average

temperature, scores of 2, 3, and 4 represented varying degrees of attention to variability. Some of the 2's may not have reasoned using variability, but recognized averages alone were insufficient. At pre-assessment 19/56 (32%) of teachers attended to issues of variability, but only 4/56 (7%) did so fully. At post-assessment, 33/56 (59%) of teachers attended to issues of variability and 18/56 (32%) did so sufficiently to score 4. On the post-assessment, seven teachers scoring 2, indicated that they would like to run a randomization test on the distributions. These responses suggest that some teachers may have over-generalized the utility of the procedure. Similarly, one teacher suggested a sampling approach that was not entirely inappropriate, but may have been an artifact of the professional development activities.

Figure 20 displays the distribution of pre-assessment, post-assessment, and gain-scores for Item 2. It also includes a cross-tabulated display and summary of the ways in which teachers' scores changed from pre to post on this item. The disposition to think of variability in conjunction with averages when comparing distributions can be seen to have markedly changed based on responses to Item 2. The change was not perfectly positive as can be seen in a sample of responses in Table 27. Many teachers demonstrated dramatic changes in their use of statistical language and attention to both center and variability in their responses.

The most common score on Item 2 changed from 1 (27/56 [48%]) on the pre-assessment to 4 (18/56 [32%]) on the post-assessment. Thirty-one of 56 (55%) teachers improved their scores on this item from pre- to post. Gain-scores were significantly greater than 0 and the final average score on this item hovered around 2.66, up from 1.82 (effect size 0.60) at the pre-assessment. Improved responses indicate sensitivity to

standard deviation and normal distributions, both important and recurring aspects of the professional development intervention. Results on this item suggest teachers' improved appreciation and understanding of variability when comparing distributions.

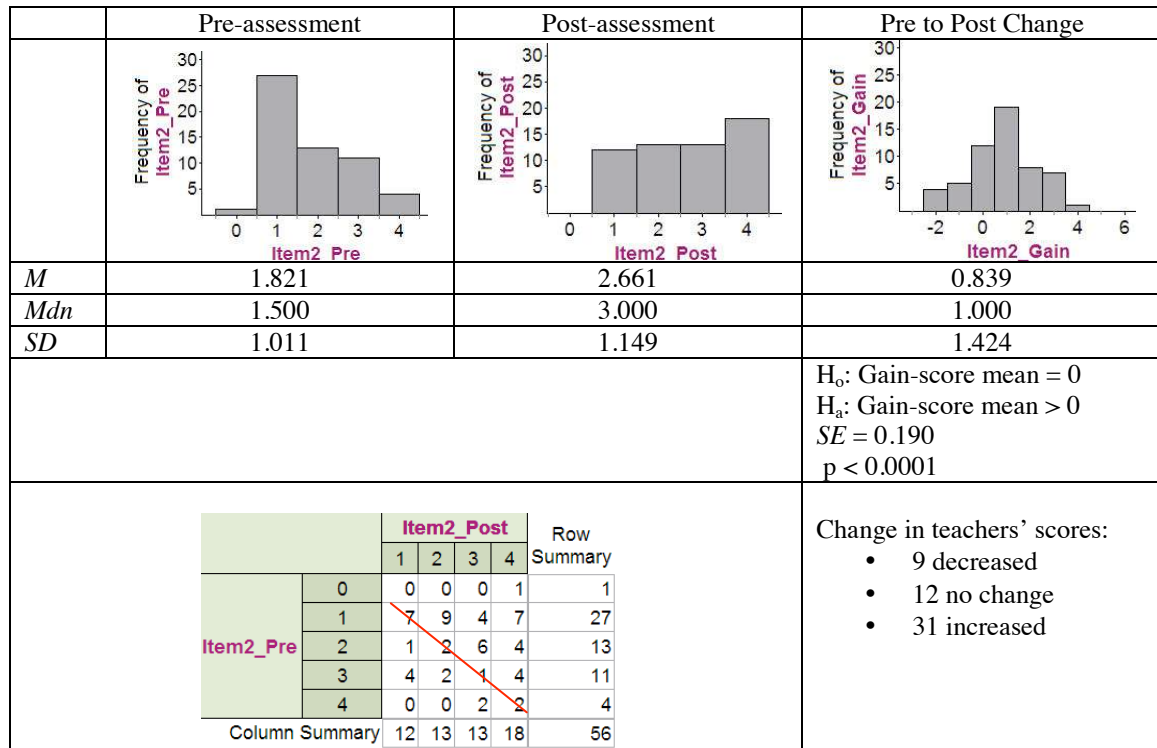


Figure 20. Item 2: Pre-, post-, and gain-score comparisons.

Analysis of Item 3

Item 3 corresponds to Task 3 on the pre- and post-assessments. Item 3, the Hospital Problem (Kahneman, Slovic, & Tversky, 1982; Makar, 2004; Shaughnessy, 1992) is a commonly seen task in statistics and probability research literature. This item was designed to elicit responses indicating reasoning on the basis of sample size. The idea is that smaller samples are more variable on average and will tend to produce proportions or sample means that may vary more from the theoretical expected value. Items of this type essentially assess understanding of sampling variability of sample

Table 27

Item 2: Sample Responses and Scores

Code #	Pre	Score	Post	Score
20	<i>I would see how a 90 degree day relates to each cities average</i>	1	<i>If you take data of the daily temperatures in the winter for each city and calculate the population standard deviation for each city, the city with the smaller pop standard deviation would be less likely to have variance in the temperature. Therefore, depending on how far a 5 degree winter day is from each cities average you could see which 5 degree day is more unusual. The more standard deviations away from the mean the more unlikely.</i>	4
22	<i>I would figure out the difference between the average and the 90 degree to see which had the greater difference.</i>	1	<i>Given only the average winter temp I would find the difference between the average and 5 to see which difference was larger</i>	1
45	<i>no clue</i>	0	<i>If both of the cities had approximately similar standard deviations and temperature had a pretty normal distribution, I could plot means and use standard deviation ideas (68% within one, 95% within two, and 99.7% within three to determine which one was more likely to have a 5 degree winter day</i>	4
51	<i>City P-->It is possible to have a few extremely ht days & a lot of mild/warm (80s)--yet have a 90 degree average). City Q-->It is also possible to <u>constantly</u> have temps in the 90's. It would be more unlikely to have a 90 degree day in city P even though the average may appear high (90 degrees)</i>	3	<i>simply given averages, I would determine which cities average temp was further away from the 5 degree day in question. Ex. X-average is 10 degrees; Y-average is 20 degrees-->it would be more unusual to have a 5 degree winter day here because of a bigger diff between 5 degrees and the average.</i>	1

proportions or sample means and frequently are seen as difficult for students to answer correctly. The pre-assessment used the actual Hospital Problem and the post-assessment used another task designed to assess similar reasoning based on the work of Well, Pollatsek, and Boyce (1990) and cited by Chance, delMas, and Garfield (2004) (see Figure 21) .

Pre-assessment Question 3

A certain town has two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50%, sometimes lower. For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days?

- A) The larger hospital
 - B) The smaller hospital
 - C) About the same number of days (within 5% of each other)
 - D) Can't tell
- Correct answer B.

Post-assessment Question 3

American males must register at a local post office when they turn 18. In addition to other information, the height of each male is obtained. The national average height for 18-year-old males is 69 inches (5 ft. 9 in.). Every day for one year, 5 men registered at a small post office and about 50 men registered at a large post office. At the end of each day, a clerk at each post office computed and recorded the average height of the men who registered there that day.

Which of the following predictions would you make regarding the number of days on which the average height for the day was more than 71 inches (5 ft. 11 in.)?

- a) The number of days on which the average heights were over 71 inches would be greater for the small post office than for the large post office.
- b) The number of days on which the average heights were over 71 inches would be greater for the large post office than for the small post office.
- c) There is no basis for predicting which post office would have the greater number of days.

Correct answer A.

Figure 21. Item 3: Pre- and post-assessment task prompts.

Figure 22 displays the distribution of pre-assessment, post-assessment, and gain-scores for Item 3. It also includes the cross-tabulated display and summary of the ways in which teachers' scores changed from pre- to post on this item. The mean pre-assessment score on this item was 2.34, making it the second easiest item on the instrument. Since scores on this item were either 0 (if blank), 1 (if incorrect), or 4 (if correct), the mean score has limited interpretation other than to signify that slightly more teachers were wrong than right. Because of the multiple-choice nature of the question and the fact that explanations were not asked for on the pre-assessment, scoring was limited to these

extreme values. Confidence in results of this question by itself is limited because guessing cannot be ruled out. On the other hand, when comparing pre to post, given the same conditions and constraints, the marked difference in distributions is suggestive of a change in understanding of the impact of sample size on the distribution of sample means.

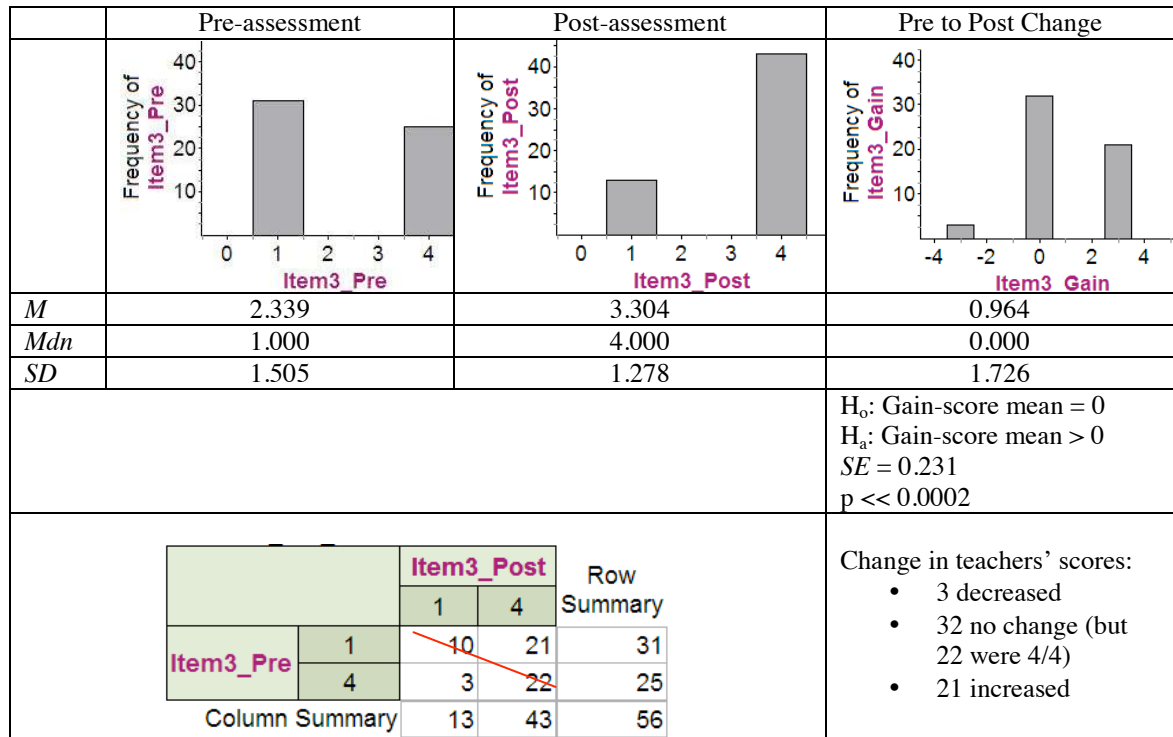


Figure 22. Item 3: Pre-, post-, and gain-score comparisons.

When the pattern of responses is examined in Tables 28 and 29, with correct responses in bold, only seven teachers answered incorrectly in the same way on both assessments.

Because explanations were required on the post-assessment, the explanations for the 13 people with incorrect answers were examined. Patterns of explanation in Table 30 suggest that some teachers were still struggling with coordinating the issue of sample size and sampling distribution of the mean.

Table 28

Item 3: Frequency of Responses on Pre- and Post-Assessment

Choice	Pre-assessment	Post-assessment
A	4—larger hospital	43—smaller post office
B	25—smaller hospital	4—larger post office
C	19—same	9—no basis for predicting
D	8—can't tell	
Total	56	56

Table 29

Item 3: Change in Responses Pre- to Post-Assessment

	Q3 Post-assessment			
Q3 Pre-assessment	A	B	C	Grand Total
A	2	1*	1	4
B	22	1	2	25
C	13	1	5**	19
D	6	1	1**	8
Grand Total	43	4	9	56

*Chose larger hospital at pre and post (1/56)

**Chose same or can't tell at pre and post (6/56)

Table 30

Incorrect Answers and Explanations for Post-Assessment Item 3

Post-assessment choice	Explanation	Frequency
B	Sample size/sample mean misconception	2
	Explanation suggests choice A	1
	No explanation	1
C	Randomization test	1
	Sample size/sample mean misconception	4
	Explanation suggests choice A	2
	Doesn't think 71 inches is unusual	1
	Need more information	1

Two examples illustrate this difficulty.

B is my choice because the large post office would have a larger sample. By chance 71 inches may happen more often. Under a normal distribution my range would be smaller in the larger sample (sketches a normal-looking distribution centered at 69) (T22).

I chose "C" because of the size of the sample. At the small post office 5 x 365 days = 1825 heights. Since this sample size is large enough to reflect the population. . . . I believe the standard error would be small so no sig. diff. could be determined (T30).

In both of these examples, teachers appear to have developed some understanding of relationship between sample size and variability of the sampling distribution of sample means. Their responses appear to indicate that they have not fully coordinated the concepts of sample, sampling distribution, and standard error, even though they are partially correct in what they say. These examples, in addition to the other responses, are reminders of the difficulty of trying to judge what another person understands. The scoring on this item was either 1 or 4. Both of these teachers scored a 1. It is fair to say that the estimates of teachers' understanding are conservatively reported. This conservatism is consistent throughout the analysis.

It is possible that the post-assessment item was somehow easier for teachers than the pre-assessment because of the contextual difference and the sampling distribution of sample means compared to sample proportions; however because the underlying mathematical idea being assessed was the issue of the variability of the mean or proportion of small samples, it is likely that the improvement in scores is due to an increased understanding of the concept rather than a change in the context. Also given that the item was the second easiest question for teachers on the pre-assessment with a mean of 2.34 and the easiest question for them on the post-assessment with a mean of 3.30 (effect size 0.56), it is likely that the professional development experience had some influence on their pattern of response. Furthermore, session activities likely to have supported a positive change might be Activity 2.4 with the random rectangles or Activity 3.4 through the investigation of the sampling distribution of sample means.

Analysis of Item 4

Item 4 corresponds to Tasks 4a, 4b, 5, and 7 on the pre- and post-assessments.

Item 4 represents a cluster of multiple choice questions designed to assess teachers' understanding of distributions. Collectively, the item assessed teachers' knowledge of the relationship between the mean and median in skewed distributions, whether they were familiar with language of skewed left or skewed right, and their familiarity with the effect of outliers on measures of center and variability. There were four questions in this cluster and the item was scored based on the number of correct responses. Figure 23 displays the distribution of pre-assessment, post-assessment, and gain-scores for Item 4. It also includes a cross-tabulated display and summary of the ways in which teachers' scores changed from pre to post on this item.

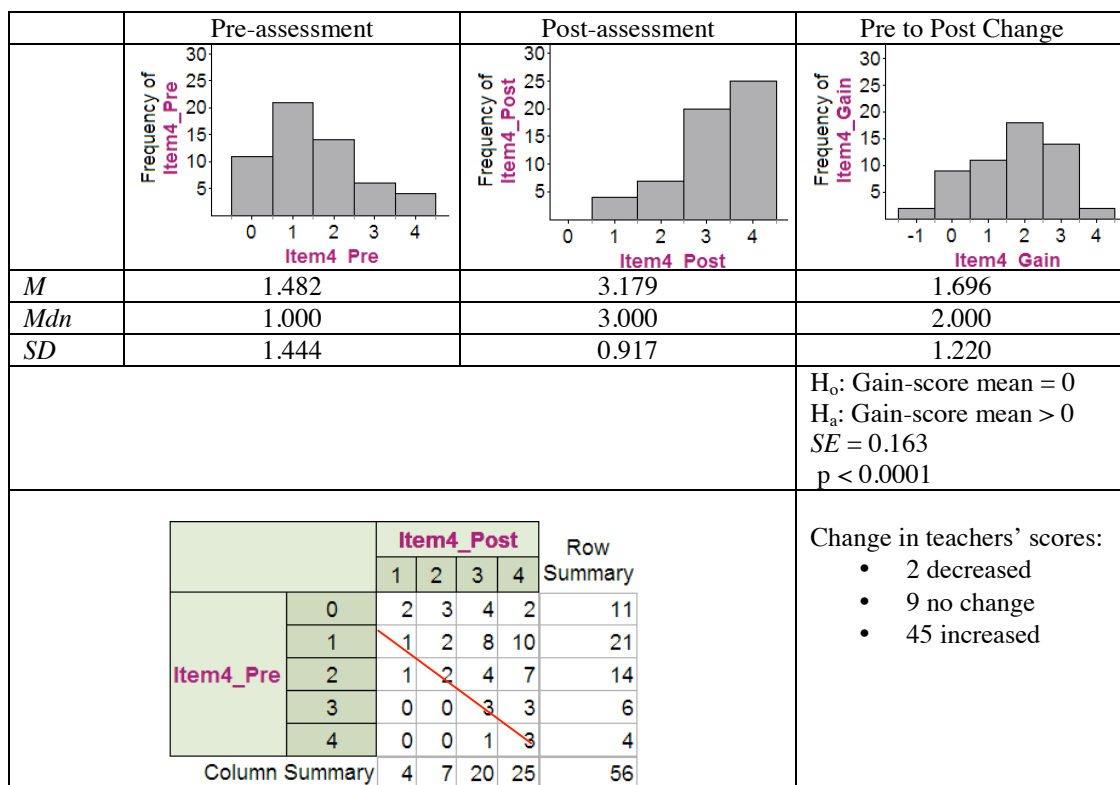


Figure 23. Item 4: Pre-, post-, and gain score comparisons.

Surprisingly, this item produced the second lowest mean score for teachers in this study at pre-assessment, suggesting lack of familiarity with some basic language of statistics. During the pre-interviews, teachers confirmed they knew how to compute mean, median, and mode, but most (all but 1) confessed to only a procedural understanding of these concepts and had little idea of the ways in which they may be used to summarize or interpret distributions and very limited knowledge of how mean and median relate to each other under various scenarios. As was seen in Chapter IV, the knowledge of the direction of skewness of a distribution was almost completely absent from this sample of teachers. Arguably, knowledge of the direction of skewness is not essential for profound statistical understanding; however, given that of 56 teachers, only 14 (25%) answered this question correctly at the pre-assessment and nearly all of the teachers (95%) answered correctly at post-assessment, it does suggest a limited exposure to statistical language which may be representative of a more general pattern of limited exposure to statistical thinking and reasoning. Table 31 contains the pattern of responses to the four separate questions associated with this item, with correct responses in bold.

All 4 teachers scoring 1 on the post-assessment answered only 4a correctly. Of the seven teachers scoring 2 on the posttest, four answered 4a and 4b correctly, one answered 4a and 5 correctly, one answered 4a and 7 correctly, and one answered 5 and 7 correctly. Of the 20 teachers scoring 3 on the pos-assessment, two missed 4a, seven missed 4b, four missed 5, and seven missed 7. Twenty-five (45%) teachers scored 4 on this item at post-assessment, up from four (7%) at pre-assessment. This item was the second most difficult for teachers on the pre-assessment and the second easiest problem on the post-assessment with a change in mean from 1.48 to 3.18 (effect size 1.39) with gain-score significantly

Table 31

Item 4: Response Patterns Pre- to Post-Assessment

Question	4a	4b	5	7
Focus	<i>Skewness</i>	<i>Median vs. mean in skewed dist.</i>	<i>Number of data values above mean in skewed distribution</i>	<i>Preferred measure of center & spread when outliers are present</i>
Pre-assessment	35 true 14 false 6 can't tell 1 blank	17 true 22 false 16 can't tell 1 blank	3 A Exactly $\frac{1}{2}$ 18 B more than $\frac{1}{2}$ above 17 C less than $\frac{1}{2}$ 14 D can't answer w/o calc. 4 blank	8 A mean & sd 2 B mean & var 0 C mean & range 5 D med & range 34 E med & IQR 5 F can't tell 2 blank
Post-assessment	3 true 53 false 0 can't tell 0 blank	42 true 9 false 5 can't tell 0 blank	6 A Exactly $\frac{1}{2}$ 43 B more than $\frac{1}{2}$ above 2 C less than $\frac{1}{2}$ 4 D can't answer w/o calc. 1 blank	0 A mean & v 7 B mean & sd 40 med & IQR 4 D med & range 0 E mean & range 2 F can't tell 3 multiple

greater than 0 and 45 teachers improving their score. It is likely that the matching plots to variables, matching statistics to plots, and investigation of measures of variability activities on Day 1 in conjunction with application of these ideas throughout the four days of professional development, were responsible for the significant pre to post change on this item. The use of *Fathom2* for investigating the change in median and mean in relation to each other when dynamically dragging data values and the use of *CPMP-Tools'* balancing histogram and estimating standard deviation features are likely technology supports responsible for teachers increased understanding of the relationships involved in this item.

Analysis of Item 5

Item 5 corresponds to Task 6 on the pre- and post-assessments. Item 5 on the assessments was designed to assess teachers' reasoning with boxplots and numerical

summaries. The task had been developed and used by Makar (2004) in her dissertation. She credits Pfannkuch and Brown (1996) with the idea for this task and designed the task to assess what kinds of information teachers would pay attention to when comparing the performance of two unequal-sized groups of students using data presented in both graphical and summary form (Makar, 2004). The item required teachers to provide three statements that would complete the following sentence: “By comparing the performance of Hispanic students with the performance of African-American students, I would draw the following conclusions . . .” Her 5-level scoring rubric for this task was used in the present study and is available with all of the rubrics in Appendix F.

As in Makar’s study, only the top two of three responses contributed to teachers’ scores on this task, making the possible scores from 0 to 10. To convert from a 5-level, 10-point scoring system to the four-level system in this study, a linear transformation of $y = 0.5x - 1$, where x is the 10-point possible score and y is the four-level score, was applied to teachers’ raw scores on this item.

In Makar’s study with 18 undergraduate pre-service teachers, 62% of responses at the pretest rated a Level 2 or lower and 21% of the responses were Level 4 or higher. In the present study with experienced teachers, 26% of responses rated Level 2 or lower while 36% of responses rated Level 4 or higher at the pre-assessment. At the post-assessment, 23% of Makar’s students’ responses were Level 2 or lower, 47% were Level 4 or higher; in the present study, 11% were Level 2 or lower, 61% were Level 4 or higher. At post-assessment, 53/56 (95%) of teachers in this study provided at least one response at Levels 4 or 5 compared to only 37/56 (66%) at the pre-assessment. Perhaps it is not surprising that experienced teachers scored better on this item at both pre- and post-

assessment than did pre-service teachers as undergraduate students. It is likely that practicing teachers have more experience with reading, synthesizing, and interpreting data by virtue of their teaching practice.

This item began as the easiest of the 10 items with a mean score of 2.47. By the post-assessment it had been replaced by three other items and ended as the fourth easiest item for teachers with a mean of 3.03 (effect size 0.53). Even with the relative ease of this item for teachers, gains significantly greater than 0 were seen pre to post. Teachers' responses indicated a generally greater disposition toward a distributional view of the data and awareness of variability following the professional development experience. Figure 24 displays the distribution of pre-assessment, post-assessment, and gain-scores for Item 5. It also includes a cross-tabulated display and summary of the ways in which teachers' scores changed from pre to post on this item.

Analysis of Item 6

Item 6 corresponds to Task 8 on the pre- and post-assessments. The intent of Item 6 was to assess teachers' understanding of simulations. At the time of the pre-assessment, it was unlikely that teachers would have been aware of or had the opportunity to study randomization distributions or permutation tests. For that reason, it was not appropriate to assess understanding of teachers' interpretation of randomization test results on the pre-assessment; however, this was important on the post-assessment because of the professional development intervention. Because of the simulation nature of the randomization test using technology, it was decided that a proxy would be to assess understanding of a simulation scenario using Monte Carlo approaches, using a task similar to the famous "Cereal Box Problem" (Travers, 1981; Wilkins, 1999). Though not

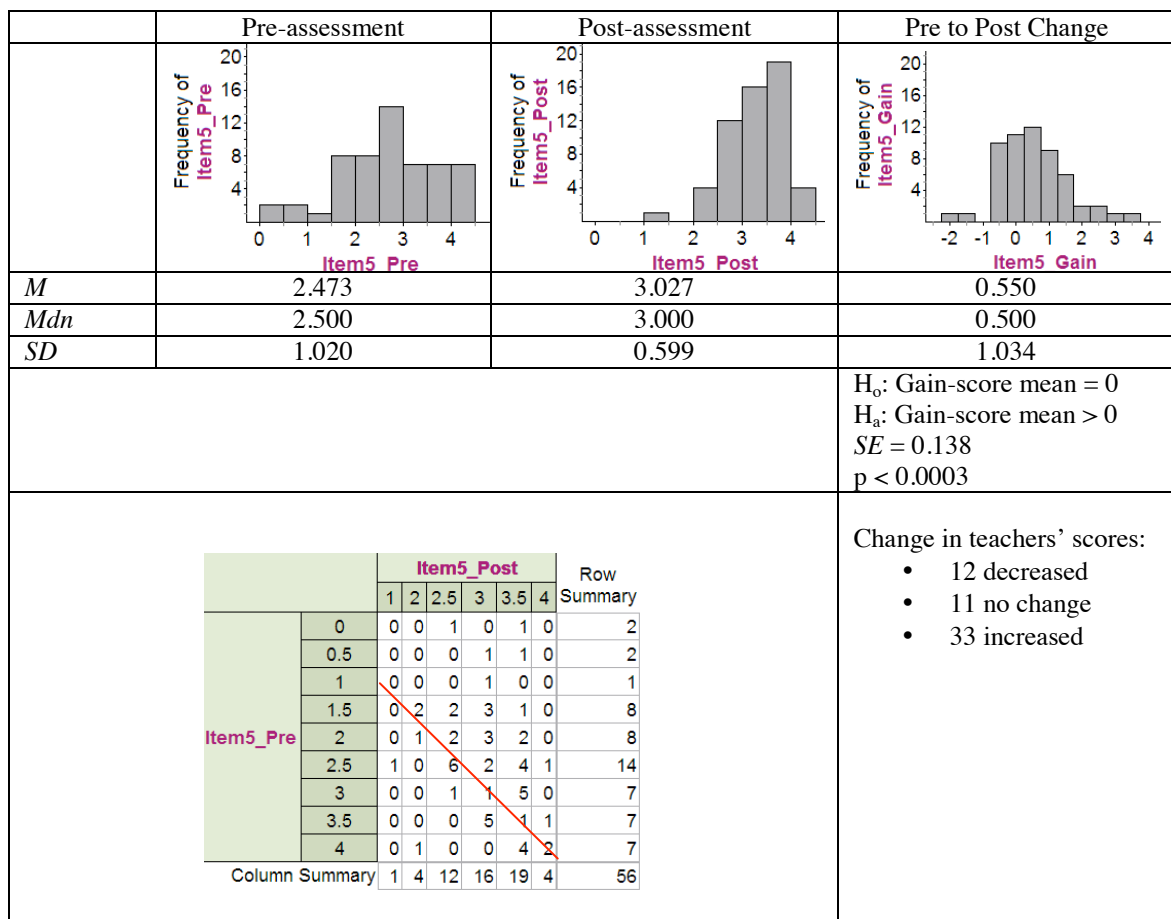


Figure 24. Item 5: Pre-, post-, and gain-score comparisons.

a perfect match, the item provided a vehicle for beginning to understand teachers' understanding of the use of simulations and experimental probability.

Table 32 provides a comparison of the general similarities and differences between the Monte Carlo process and the randomization test approach. Clearly there are differences between the procedures; however, it was expected that teachers may have some familiarity with Monte Carlo simulations prior to the professional development whereas they may never have learned about the randomization test process until the professional development took place.

Table 32

Comparison Between Monte Carlo Simulation and Randomization Testing Procedures

Monte Carlo	Randomization Test
1. Identify the model	1. Examine cases
2. Define a trial	
3. Record the statistic of interest	2. Determine statistic of interest
4. Repeat trial	3. Resample and collect many measures of the statistic of interest
	4. Create sampling distribution
5. Find an average	5. Determine p -value
	6. Interpret the result

Figure 25 displays the distribution of pre-assessment, post-assessment, and gain-scores for Item 6. It also includes a cross-tabulated display and summary of the ways in which teachers' scores changed from pre to post on this item. This item was the fourth most difficult item for teachers on the pre-assessment and the third easiest problem on the post-assessment with a change in mean from 1.79 to 3.14 (effect size 1.16) with gain-score significantly greater than 0 and 43 teachers improving their score.

Sample pre-assessment responses are provided in Table 33. Twenty-three of 56 (41%) teachers responded similarly to Teacher 02, and 17 of 56 (30%) responded similarly to Teacher 16. Forty-four of 56 (79%) the teachers' responses to this task scored at Level 2 or below, suggesting a deterministic approach to interpreting simulation results. These teachers indicated no problem with the small number of trials or the variability in the data and did not attempt to reason beyond the data. The remaining 12 teachers indicated at least a consideration of variation, but only one of the 12 scored a 4 on this item.

The randomization test task on the post-assessment was a simulation-type of problem and presented teachers with original experimental data, summary statistics, side-by-side boxplots and dotplots of the data, and a simulation-based randomization

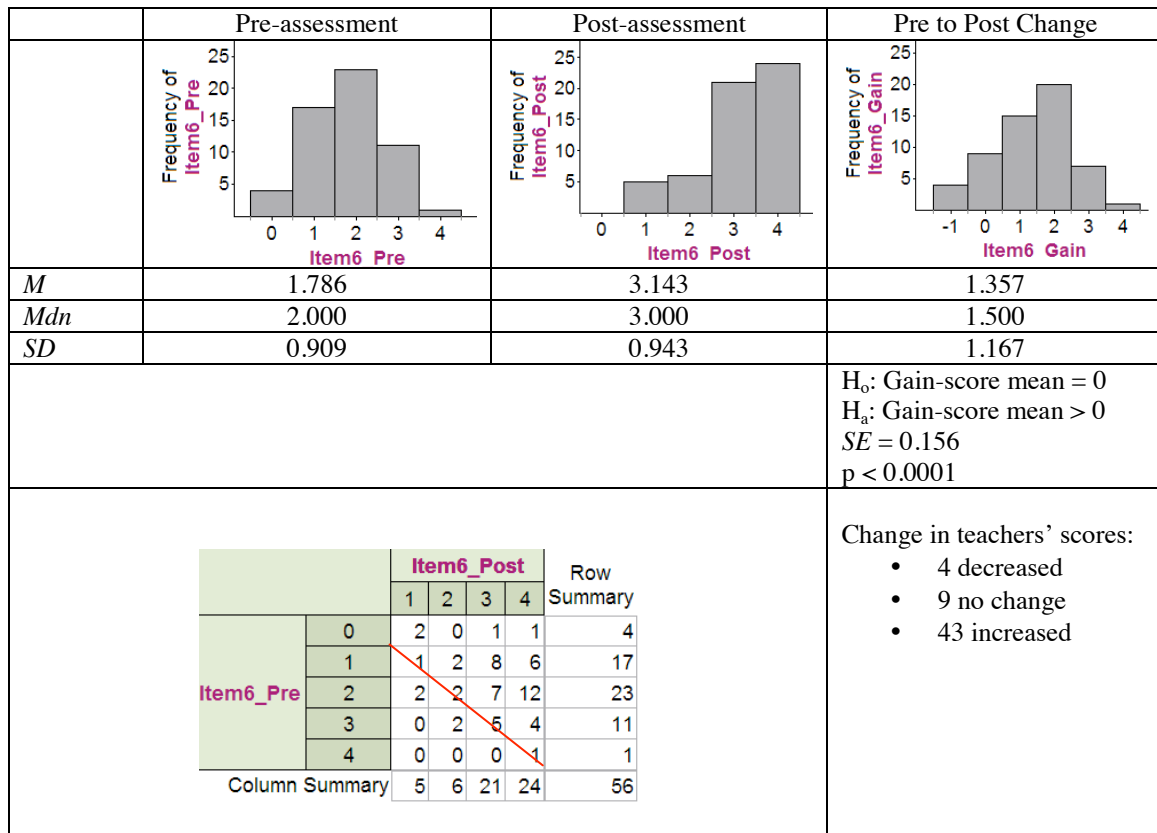


Figure 25. Item 6: Pre-, post-, and gain-score comparisons.

Table 33

Item 6: Sample Pre-Assessment Responses

Score	Sample Response
0	<i>I don't know (T62).</i>
1	<i>Somewhere between 10 and 21 boxes as this the max/min of boxes required in trials (T16).</i>
2	<i>average=15.6 Theoretically , they should have to buy 16 boxes because that is the mean of the data (T02).</i>
3	<i>With a mean of about 16 and a sigma approx 4, he has a fair chance of getting all 6 with a purchase of 16 and should expect to get all 6 prizes with a purchase of 20 (T10).</i>
4	<i>Keegan really needs to do a much large # of trials. By only doing 5 trials he really can't tell whether 10 and 12 are realistic # of boxes or are 19 & 21 more acceptable. But, based on Keegan's simulation, he has a mean of 15.6 and standard deviation of 4.615 and a box plot of (shows box plot here). Keegan can expect to buy at least 16 boxes in order to have an approx. 50% chance of getting all 6 racers. He may need to purchase as few as 10 or 11 or as many as 20 or 21 though--b/c his # trials is so small his variability is also going to be quite high (T35).</i>

distribution from which to reason. Forty-five of 56 teachers (80%) scored a 3 or 4 on this item at the time of the post-assessment, indicating a strong understanding of interpreting the analysis using the randomization distribution. To score 4 on this item, teachers needed a concluding statement in the context of the problem; those scoring 3 had a correct interpretation but failed to connect the problem context to the analysis. Examples of Level 3 and 4 responses are below.

Level 3

diff in means = 0.82. It would appear that the likelihood of the same results of the experiment occurring strictly by chance is minimal. We can conclude that there is a significant difference in the data [marks .82 on the RD] (T59).

The differences in the means between treatment/control = .82. This is shown as marked in the Randomization distribution. Because the P-value appears to be less than .02 the probability of getting results randomly above this mean difference (all factors being equal) is small and the original difference of .82 is statistically significant. [marks .82 on the distribution and indicates p-value by circling extreme values; makes a note at the prompt: did they (control group) receive any treatment?] (T16).

Level 4

diff .82. The p-value would be low (there are not many values located above the observed diff, in the original means) Therefore, there is strong evidence to suggest that the diff. in mean in the original data is NOT due to chance, but is in fact due to the treatment given (the drug). Therefore, Drug A prevents low birthweight (according to this study) [makes a mark on the graph at .82 and circles the extreme values] (T51).

I would conclude, based on the box-plot and randomization test, that the test is valid and that drug A does prevent low birthweight. The p-value is less than 0.05 (I'm assuming by looking at info) and the boxplots don't really overlap). [there is a mark on the graph that may represent .82] (T36).

Six teachers scoring 2 all indicated that the difference in the treatment means was significant, but their explanations were insufficient to score more highly. Four of the

remaining five teachers scoring one on this item used a variety of incorrect reasoning. The fifth teacher indicated that the difference was significant but couldn't explain why. As a check of whether prior statistics coursework may have been associated with a low-level response, Table 34 confirms that scores of 1 were received by teachers having taken 0, 2, or an undisclosed number of prior statistics courses. More importantly, the proportion of teachers scoring 3 or 4 by number of statistics courses taken is strikingly high, regardless of prior course experience.

Table 34

Item 6: Comparison of Performance on Post-Assessment Item 6 with Number of Statistics Courses Taken

		Score on Post-assessment Question 8				Row Summary		Proportion of Levels 3 or 4
		1	2	3	4			
Number of Statistics Courses Taken	No response	1	0	1	1	3		2/3 (67%)
	0	2	0	3	2	7		5/7 (71%)
	1	0	5	13	12	30		25/30 (83%)
	2	2	1	3	8	14		11/14 (79%)
	3	0	0	1	1	2		2/2 (100%)
Column Summary		5	6	21	24	56		

This question on the post-assessment is particularly important for the interpretation of this study's results because it provides evidence that teachers, in a short period of time, could come to make sense of the logic of randomization testing as well as articulate that understanding in an intelligible way, regardless of prior statistics background. The logic of hypothesis testing continues to stump many, even after repeated statistics education experiences (Krauss & Wassner, 2002). Given the performance on an item requiring teachers to grasp the process as well as to explain their reasoning, this evidence may suggest that the randomization test as a pedagogical and statistical device may be worthy of further consideration and investigation.

Analysis of Item 7

Item 7 corresponds to Task 9 on the pre- and post-assessments. Item 7 posed an experimental situation with random assignment and presented the results of the treatment and control groups as side-by-side dotplots. The item was adapted from a problem in *Introduction to the Practice of Statistics*, Chapter 14 (Moore & McCabe, 2005, pp. 47–48). The content of this task was related to teaching in that students were randomly assigned to experience a new educational innovation, in this case, directed reading activities, or the traditional teaching method. Based upon the evidence provided in the graphical displays, teachers were asked to determine whether the new method was better than the traditional method, explain their decision, and describe any reservations about their conclusions. Figure 26 displays the distribution of pre-assessment, post-assessment, and gain-scores for Item 7. It also includes a cross-tabulated display and summary of the ways in which teachers' scores changed from pre to post.

This item was designed to assess teachers' understanding of both interpreting side-by-side dotplots as well as their understanding of basic experimental design. At the time of the pre-assessment, 11 of 56 (20%) teachers scored at Level 3 or 4; whereas at post-assessment, 33 of 56 (59%) teachers scored at Level 3 or 4. Most teachers scored at Level 2 on the pre-assessment and Level 3 on the post-assessment. Samples of teacher responses in Table 35 help to focus on the nature of the change in language, analysis, and understanding of comparing distributions.

Teacher 12 focused only on the low values to make a comparison on the pre-assessment; on the post-assessment this teacher was not confident to make a claim but did recognize the randomization test as a vehicle to test whether the difference in the means

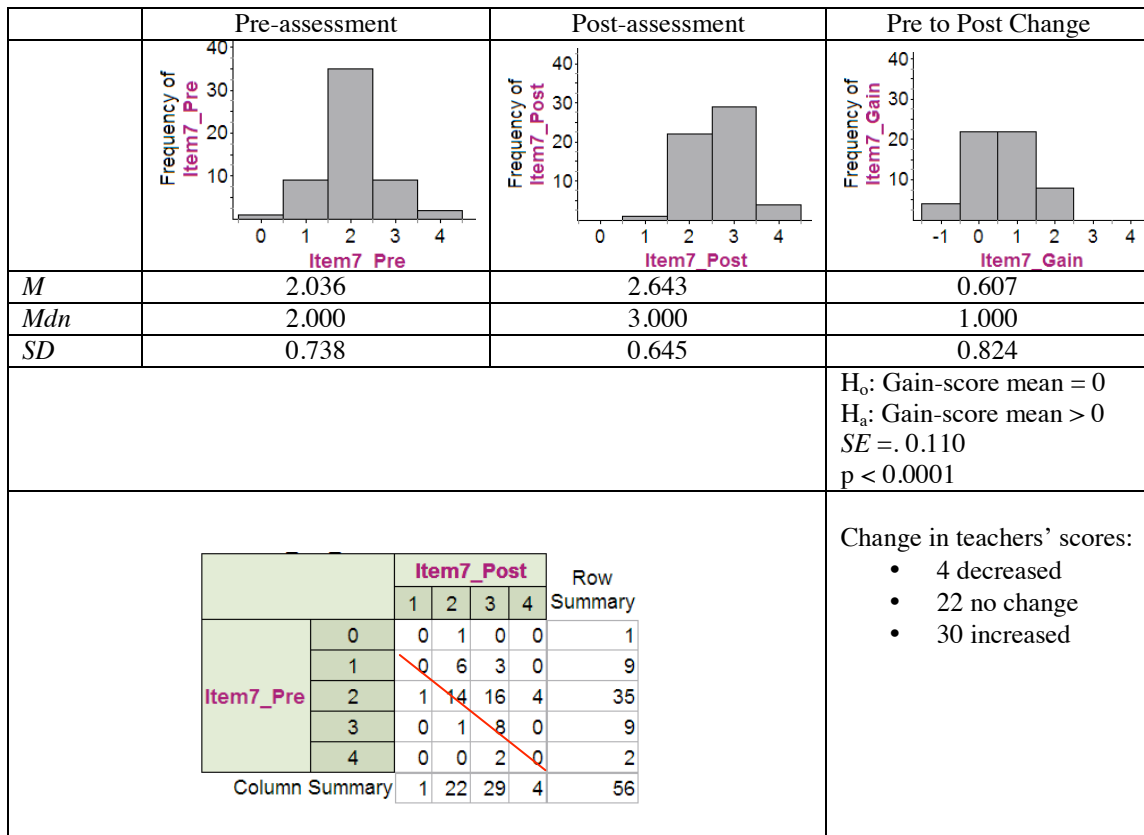


Figure 26. Item 7: Pre-, post-, and gain-score comparisons.

from the distributions was significant. Teacher 93's reasoning advanced from an apparent gut-level response to attending to spread and outliers in the distributions and again recognizing the randomization test as an appropriate tool. Teacher 07 first reasoned from a middle cluster approach augmented with what seemed to be a dose of teacher-biased explanation and later mentioned the randomization test, overlap in the dotplots, mean in relation to standard deviation and a faulty conclusion. Teacher 02's post-assessment response suggests improved facility with statistical language of measures of center and spread and some understanding of differences due to chance or treatment. Teacher 36 appears to have an understanding of measures of center and spread at the pre-assessment but fails to fully address the reservation aspect of the question. At the post-assessment,

Table 35

Item 7: Sample Responses and Scores

Code #	Pre-assessment	Score	Post-assessment	Score
12	<i>From the control you can compare results and note that there are fewer results in the 20s to 30s in the treatment from the control, so something helped. [circled values from the control group on the graph and drew an arrow to the treatment group]</i>	1	<i>I can't really say if there is a difference. I would have to run a R.T. Test to see if the mean difference was significant.</i>	2
93	<i>I don't know, but my instinct says the treatment group is better.</i>	1	<i>At first glance, the treatment group appears better than the control group. The treatment group scores are not as spread out as the control group scores. I would do a randomization test to see if the difference in means was statistically significant. I would also look at the randomization of the standard deviations as well, out of concern for the outliers in the data.</i>	3
07	<i>From this data it would appear that the traditional method is not working for those 6 or so students that have the lowest scores. The new method seems to have kept the majority of the student achieving at a level between 42 & 62. While the top group is clustered between 36 & 56 with far more lagging behind. The one exceptional student from the top group is probably not attributed to the method, but own ability to succeed. The New method shows great merit for greater success for all w/fewer "left-behind!" No reservations. [indicated clusters on the graphs]</i>	2	<i>I would have to perform the randomization test & see where the difference of the 2 actual means falls in relation to 2 std deviations either way of the mean for the test. My guess is that it would not be significant since from this dot plot it looks like the majority of the dots overlap. So my conclusion is that the new test does not show significant improvement. [drew normal-looking curves over each distribution and marked off the middle section of each set]</i>	2
02	<i>In the treatment group fewer students scored under 40%, based on the graphs. However we do not know how much they <u>improved</u> by, that group may have already scored that high. They should have compared before and after scores.</i>	2	<i>The treatment group scores appear to have a higher mean and median, with a smaller variance. However there are very few data points, and without a before/after comparison we have no way to know if these results are "by chance or by treatment".</i>	3
36	<i>The median is higher for the treatment group, there is less deviation for the treatment group. I would also calculate the mean, standard deviation. [located the medians on the graphs, lists sample sizes]</i>	2	<i>[Drew boxplots over the existing dotplots] The median (and I would assume) the mean is higher for the treatment group. There is less variance in the treatment group as well. Because of the overlap of the boxplot I would be hesitant to draw any conclusion without performing a randomization test.</i>	3

Table 35—Continued

Code #	Pre-assessment	Score	Post-assessment	Score
39	<i>Based on a brief glance, it appears the # of students who have lower scores after the treatment has declined. I would calculate mean & st. dev. My reservation is the small sample size [underlined].</i>	2	<i>[marked mean of 42 control; 52 treatment] Since the mean of the treatment group is much higher than the mean of traditional method, this may be evidence to support the treatment. The range of control group is approx 75 & range of treatment group is approx 49. Since the treatment group has higher mean w/ less variability, this is evidence that the new method works. To determine if this difference is stat. sign., I would perform a randomization test to see what % chance this could have occurred by random chance. I would have reservations about the small sample size of this study.</i>	4
97	<i>Evidence that the new method is better [constructed boxplots on top of the dot plots]: --75% of the students scored at/above 24 in the control group. With the new method, 100% scored at/above 24. --50% of the students in the control group scored at/above 44. 75% of the new group did. --25% of the control group was at/above 54 while 50% of the new group did. Reservations: Did all kids improve? did some stay the same or do worse?</i>	3	<i>Because there are fewer scores below 42 in the treatment group, I might conclude that the treatment helped improve scores. However, with so few samples, I may not have the whole picture. Also, the difference may only be due to chance. So more samples & a randomization test would help.</i>	3
23	<i>Evidence: 1) range is smaller (lower end has been brought up). 2) mean/median are higher. Reservations: 1) human error. 2) only 1 study, may need more to produce accurate results. 3) Accuracy of DRP. 4) Is a median DRP value of 53 good or should a new method be studied.</i>	3	<i>could have used their previous DRP Scores to help us look at the new scores. The Treatment group did slightly better on average and since we always need a higher reading score I would have to say the treatment is certainly worthwhile [made marks at what may be the medians of each set on the graphs]</i>	2

Teacher 36 provided a similar response but appeared skeptical about making a conclusion without results of a randomization test. Teacher 39's responses evolved from an additive comparison of number of low student scores and suggestion of calculating mean and standard deviation to estimating the mean of both groups, discussing differences in variability between groups and suggesting use of the randomization test. Teacher 97

scored at Level 3 on both assessments, but the responses progressed from simply comparing quartiles toward sensitivity for the need for larger samples and randomization testing to be confident in conclusions. Finally, Teacher 23 represents a decline from pre- to post-assessment. At the pre-assessment, this teacher addressed measures of center, spread, and plausible reservations about any conclusion. At the post-assessment, Teacher 23's response is less thorough than the pre-assessment response and failed to address issues of spread or any real specific detail.

Analytically, it is clear that experiences with the randomization test influenced teachers' responses to this question. It may have even prevented some teachers from comparing some features of the distributions such as shape, center, and spread. On the other hand, it is equally clear from the analyses of the responses that a greater percentage of teachers were sensitive to measures of center and spread and able to produce statements and justifications from a more statistical perspective. Some of the skepticism in making the comparison at the post-assessment may have been an artifact from Activity 4.2, The Physician's Health Study, described in Chapter IV.

Figure 27 contains sample statistics and boxplots of the distributions from Item 7.

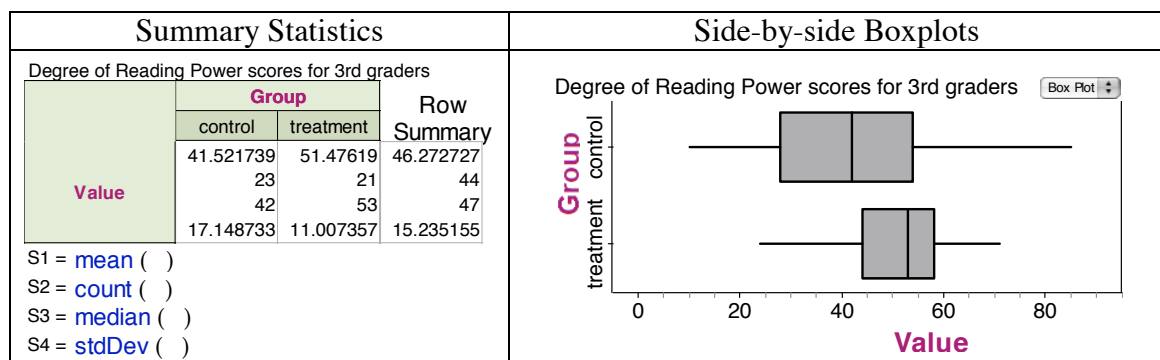


Figure 27. Item 7: Summary statistics and side-by-side boxplots from the task.

Under the assumption or null hypothesis that the two samples from this task are from the same population, that is, there is no difference between treatment conditions, two randomization tests, each with 10,000 trials produced the following results: The probability that a mean difference of 9.96 or greater would happen by chance alone is approximately $149/10,000 = 0.0149$; the probability that a median difference of 11 or greater would happen by chance alone is approximately $319/10,000 = 0.0319$. Both of these results strongly support statistically significant differences in the means and medians between the control and treatment group and the conclusion that the treatment group performed significantly better than the control group. Student's t , using unpooled variances, is 2.311 with 37.8554 degrees of freedom. If it were true that the population mean of the treatment group were equal to that of control (the null hypothesis), and the sampling process were performed repeatedly, the probability of getting a value for Student's t this great or greater would be 0.013. Not only are the differences in means and medians significant, they are highly significant.

Teachers' responses to this question suggest that making these kinds of comparisons without tools, such as the randomization test, is cognitively demanding. Because teachers had become acquainted with the randomization test through the professional development experience and because their intuition had been challenged and in some cases even compromised when their predictions turned out to be misguided, some teachers may have developed a more conservative approach to their decision-making. Teachers' familiarity with the teaching context of this problem may have helped some teachers and hindered others.

This question had relative difficulty indices of 7 at the pre-assessment and 4 at the post-assessment and only four teachers scored at a Level 4 on the post-assessment. It went from the fourth easiest question for teachers to the fourth most difficult question. The pre-assessment mean was 2.04, post-assessment mean was 2.64 (effect size 0.74), gain scores were significantly greater than 0, and the post-assessment distribution was positively shifted from the pre-assessment distribution.

Analysis of Item 8

Item 8 corresponds to Tasks 10, 11, and 12 on the pre- and post-assessments. These tasks were collectively designed to assess teachers' reasoning about comparing variability from one distribution to another while simultaneously attempting to determine whether teachers demonstrated a consistent approach to judging variability or utilized a variety of methods. Task 10 presented teachers with two symmetric distributions as dotplots both centered at 3 with range from 1 to 5, the first with $n = 3$, the second with $n = 13$. Task 11 presented teachers with two different symmetrical distributions as histograms both centered at 20 with range from 15 to 25, with the first appearing unimodal and normal-looking while the second appeared bumpy and multi-modal. Task 12 was adapted from Makar's dissertation study (2004) in which she designed the task to determine whether teachers reasoned about variability in terms of bumpiness or range or some other heuristic. Collectively, this combination of tasks was designed to determine whether teachers' understanding of variability may evolve toward reasoning about average deviation about a central anchor point.

Scoring of this item involved a coordination of responses to all three questions (see rubric in Appendix F). Figure 28 displays the distribution of pre-assessment, post-

assessment, and gain-scores for Item 8. It also includes a cross-tabulated display and summary of the ways in which teachers' scores changed from pre to post on this item.

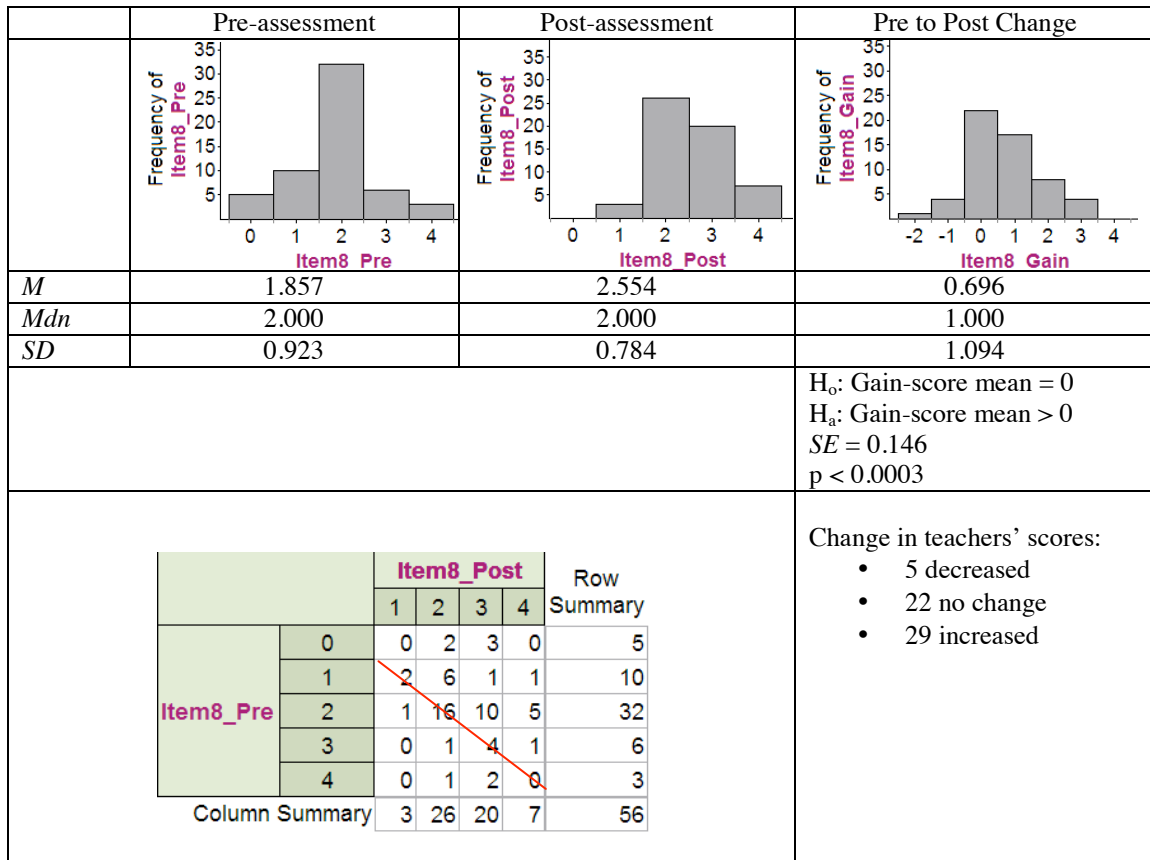


Figure 28. Item 8: Pre-, post-, and gain-score comparisons.

The relative difficulty index for this item went from a 6 at pre-assessment to a 3 at post-assessment, suggesting that it turned out to be one of the more difficult items for teachers. When the pattern of responses is examined in Table 36, positive changes are seen for Tasks 10 and 11, but Task 12 responses were approximately the same pre-to post. Sample responses to the question are presented in Table 37.

Scores on Item 8 privileged reasoning over choice of correct distribution. Teacher 03 appeared to lack a consistent or well-developed view of variability, but provided some evidence of reasoning from a normal curve perspective and invoked “bumpiness” as a

Table 36

Item 8: Question by Question Comparison Pre- to Post-Assessment

	Pre-Q10	Post-Q10	Pre-Q11	Post-Q11	Pre-Q12	Post-Q12	Pre-All 3	Post-All 3
Correct Choice	38 (68%)	50 (89%)	38 (68%)	50 (89%)	21 (38%)	19 (34%)	11 (20%)	15 (27%)
Incorrect Choice	18 (32%)	6 (11%)	18 (32%)	6 (11%)	35 (62%)	37 (66%)	45 (80%)	41 (73%)
	56	56	56	56	56	56	56	56

proxy for variability. Teacher 07 demonstrated a similar lack of understanding of variability but does select a correct choice for Task 12 based on reasoning from the range; the reasoning presented for Tasks 10 and 11 appear to suggest this teacher thinks of variability as the number of different values the variable may take on. Teacher 36 correctly identified all three graphs with the most variability, but has reasoning for Tasks 10 and 11 suggestive of a frequency approach of looking for repeat data values which may be related to central clustering, although it is not clear from the explanations given. This teacher then reasoned from the range on Task 12. Teacher 2061 reasoned from a standard deviation perspective and clearly articulated an understanding of variability as average distance from a central anchor point. The choice on Task 12 was incorrect, however, the reasoning was consistent and plausible. Teacher 40 appears to have coordinated sample size, average distance from the mean, and range in order to correctly select all three choices and provide sensible reasoning representing a statistical perspective.

Understanding how others understand variability has become a focus of recent research as seen in Chapter II. Teachers in the present study conveyed multiple ways of understanding variability when contrasting variability between two distributions which included variability as bumpiness, variability as whole range (the spread of all possible

Table 37

Item 8: Sample Responses and Scores

Code #	Q10	Q11	Q12	# of Correct Choices	Level
03-pre	<i>B--because it allows for more choices. Graph A allows only for 3 options</i>	<i>Course 1 ~ reasoning do not know.</i>	<i>School B - height ranges in students do not normally follow a bell-shaped curve. The heights can vary greatly.</i>	0	1
07-pre	<i>B has more variability, where group A had an = dist.</i>	<i>Groups have same variability w/11 different ages.</i>	<i>Group A has more variability w/hts varying from 145" to 166" & B only from 147 1/2 to 162 1/2</i>	1	2
36-pre	<i>A -- no repeats in A, for B 3 is the greatest/most common</i>	<i>CPR #2 - even though they have the same range, #1's Age are around 20.</i>	<i>A--it has a greater range of scores</i>	3	3
2061-post	<i>[Calculated SD for both--1.63, 0.96] A The data is spread out. Their distances from the mean are 2, 0, 2. For B, there are more values at or close to the mean so their differences are 2, 1, 1, 0, 0, 0, 0, 0, 1, 1, 2. When calculating sigma by hand, group B results in a lower value.</i>	<i>[Calculated SD for both, 2.3, 2.8] The data is more clustered about the mean for Course 1. For Course 2, more data pts are further away.</i>	<i>[pointed to peaks in dist. B] Same reason as for #11. The bars that are tall would have many values that are quite a distance from the mean, resulting in a higher sigma.</i>	2	4
40-pre	<i>A has more, B has less variability because n is greater and smaller % of data differs from mu [provided formula for variance]</i>	<i>CPR course #2 has more variability because a higher % of data are spread further from the mean.</i>	<i>School A appears to have more variability because the distribution is more spread but this is difficult to tell because school B appears almost bimodal so that might create a larger variance upon calculation [drew a smooth bell-shaped curve over each distribution]</i>	3	4

values), variability as range (difference between maximum value and minimum value), variability as the sum of the residuals or standard deviation, more variability meaning not normally-distributed, and more variability as less data values. Many of these are similar to those summarized by Shaughnessy (2007). From pre- to post-assessment, there was a marked increase in the number of responses suggesting attention to standard deviation or some type of distance from a central anchor point.

Teacher 17 (Table 38) represents teachers whose choices matched at pre- and post-assessment, but whose reasoning appeared to evolve toward a statistical perspective, even though initial conceptions appear to be somewhat resistant to change. Teacher 57 is an example of someone who selected all three correct choices on both assessments, but at the pre-assessment, the reasoning was devoid of a recognizable statistical perspective; whereas at the post-assessment, this teacher's statistical reasoning is dramatically improved.

One misconception that appeared not to be resolved by the post-assessment for some teachers was that normal distributions have less variability than non-normal distributions. There were a number of explanations from teachers that suggested they decided about relative variability on the basis of whether one distribution was more "normal" than another. This is somewhat challenging to reconcile given that over the course of the professional development experience, teachers explored a number of normally-shaped and non-normally shaped distributions and had the opportunity to compare normal curves with a variety of centers and spreads. The only activity that directly and intentionally involved the comparison of variability of normally-shaped distributions with non-normally-shaped distributions was Activity 1.4 Standard Deviation

Table 38

Item 8: Pre- and Post-Assessment Comparisons for Two Teachers

Code #	Task 10 Response	Task 11 Response	Task 12 Response	# Correct Choices	Level	General Reasoning
17-pre	<i>Group 1 because a larger percent of the group falls in the outliers</i>	<i>Course 2 because the frequency does not gradually increase or decrease. It goes way up and then way down.</i>	<i>School B because very few students are the mean height.</i>	2	2	extremes, bumpy, anchor
17-post	<i>There is not really enough data in group 1 to make a choice but if I had to, I would say that group 2 is normal so 2/3 of the data falls w/in 1 st. dev. of the mean and only 1/3 of the data in G1 is w/in one st. dev of the mean.</i>	<i>C2 because it seems Quintamodal [smiley face] C1 looks normal. C2 seems to have a larger st. deviation.</i>	<i>SB because the graph does not seem to follow any pattern. It would not be considered normal or bimodal. Even though SA has a larger range, I would consider SB to have more variability - the lowest # of heights are at the mean.</i>	2	3	normal, standard deviation, shape, range
57-pre	<i>Group 1--Based on percentage, a large percentage of students in Group 2 are in the middle</i>	<i>Course 2--A large percent of students are in the middle with course 1, where as w/course 2 the students' ages are about equal for each group.</i>	<i>School A--There are more heights recorded at School A.</i>	3	2	mixed, middle cluster, sample size
57-post	<i>Group 1--the std. dev. Is more. In B, a larger percentage of data is clustered in the middle near the mean.</i>	<i>Course 2-- Course 1 has a smaller std. dev., will a greater percentage of data closer to the mean. Course 1 distribution is more normal.</i>	<i>This one is tougher. School A is more normal shaped, but I think School b has a larger percentage of data clustered around the mean, so it would have less variability & a lower std. dev.</i>	3	4	standard deviation, anchor

and Its Interpretation. Part of Activity 1.4 involved teachers matching statistics (mean, median, and standard deviation) to distributions (normal, skewed, bimodal, uniform). The medians of each distribution were identical, but the means, standard deviations, and

shapes varied. Part of this activity was described in Chapter IV, but it is possible that the follow-up activities focusing on the standard deviation and its relationship to the normal distribution may have confounded teachers' understanding of variability to some degree. Because reasoning about relative variability of distributions included reasoning about normal distributions at the pre-assessment, the origins of this line of thinking is not clear. This is a finding worthy of continued investigation.

Item 8 was challenging for teachers. The scores changed from mean 1.86 to 2.55 (effect size 0.64) across assessments. Gain-scores were significantly greater than 0 for the group. Evidence from this combination of problems suggests that teachers' understanding of variability did evolve toward a more statistical perspective. Furthermore, this combination of tasks seemed to help illuminate the often numerous ways teachers may reason about variability.

Analysis of Item 9

Item 9 corresponds to Task 13 on the pre- and post-assessments. Item 9 was intended to assess teachers' ideas about the design of experiments or observational studies. The question was engineered to indirectly assess teachers' understanding of experimental design and analysis by invoking their pedagogical content knowledge as teachers. This question was designed in this way for its potential to communicate to teachers that this assessment respected their position as educators. By constructing a context in which teachers' professional experience might be privileged and their statistical knowledge highlighted simultaneously, it was hoped that this task would be a

subtle cue to teachers that the professional development they were about to participate in would be sensitive to their work as teaching professionals.

Teachers were presented with a contextual cover-story and the following directions:

Your task is not to design a study but rather to describe what you would expect a “high quality” student response to look like in this case. Please include the statistical ideas that are important to consider in this situation. (e.g., what needs to be included and addressed to be considered thorough—no need to mention neatness or organization for presentation).

It was hypothesized that teachers might expect students to design and conduct an experimental or observational study from which the opportunity to compare distributions might emerge along with potential connections to testing for significant differences via generation of sampling distributions. As the rubric in Appendix F illustrates, scoring on this item was designed sensitive to (1) formulation of a researchable question, (2) data collection, (3) data analysis, and (4) interpretation. Figure 29 displays the distribution of pre-assessment, post-assessment, and gain-scores for Item 9. It also includes a cross-tabulated display and summary of the ways in which teachers' scores changed from pre- to post on this item.

The contexts for the pre- and the post-assessment were very different and thus direct comparisons between the two sets of results were extremely challenging. On the pre-assessment, there were no scores at Level 4 and 33 of 56 (59%) scored at Level 2, suggesting some knowledge of experimental design, but either a very limited view, a failure to connect to the context of the task, or lack of enough detail to thoroughly address the task. For example, typical Level 2 responses at the time of the pre-assessment included:

- *Consider all types of environments people live in.
- *Gather large quantities of data from each environment.
- *Support w/variety of graphs. *Support conclusions based on data. (T07)

Large sample size. 2. Random sample selection. 3. List of survey questions that reflect no bias. 4. Data organized in format including labels & title. 5. Conclusions reached. 6. Collection method described (How did they make sure sample was random.) (T19)

The students would have to describe the following aspects of their study:
 Control: How are other variables besides those being studied eliminated.
 Randomization: What techniques are being used to ensure that groups of experimental units are very similar. Replication: How many experimental units are used and why. The students would also have to describe their experimental units and variables of interest in detail. (T21)

First decide on sample size & origin of sample Need to get people from farms. Use a box plot to show those who have & do not have allergies. Show median. Don't know--cannot make sense of what I'm thinking, not sure of correct sample. Need to show median, # of people in both categories, probably sort by age-older people have longer to develop immunities. (T26)

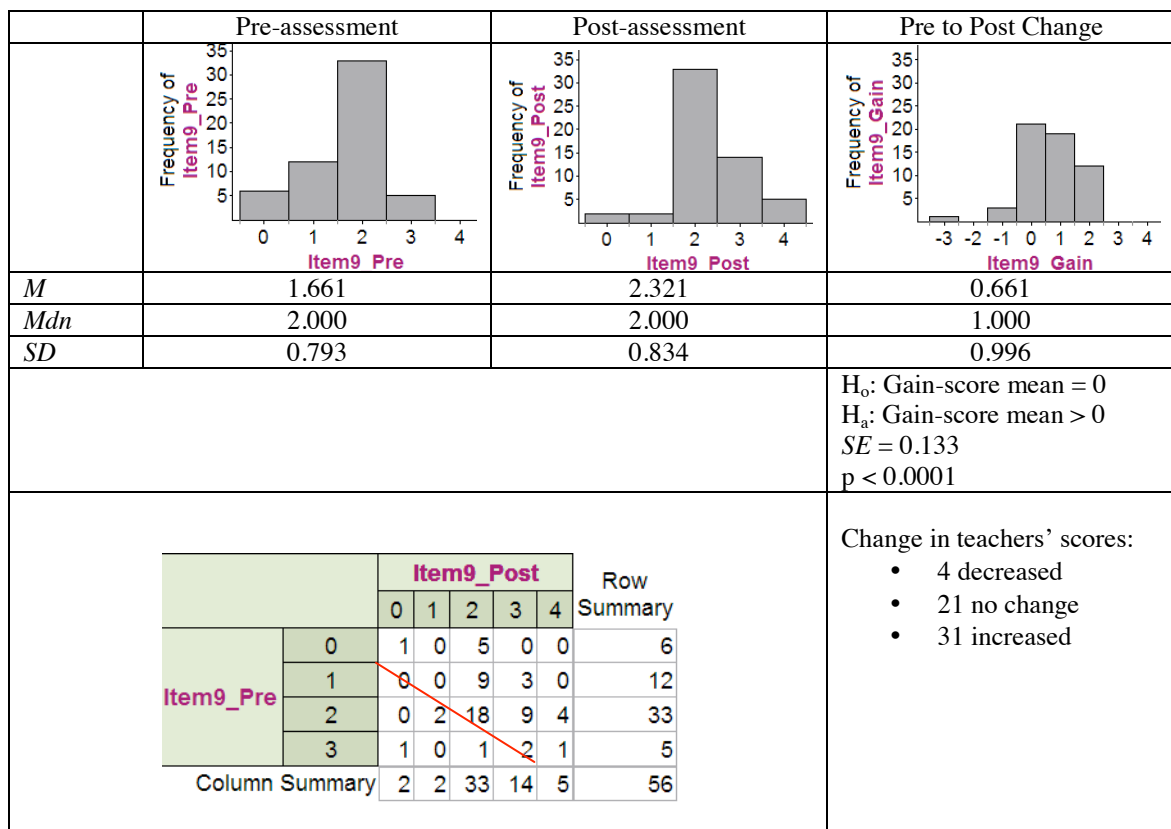


Figure 29. Item 9: Pre-, post-, and gain-score comparisons.

Attention to the formulation of a researchable question is notably absent in the preceding examples as well as in every teacher response at the pre-assessment. It is clear that teachers had some idea about the conduct of experiments; however, there was no attention to taking a given context and creating a manageable, tenable, researchable question. Assuming a question was provided, as is often the case in textbooks, many teachers appeared to have preliminary ideas about what comes next, but even so, most did not connect to the context of the given task. Instead, they wrote from what seemed to be a procedural basis.

Several examples comparing pre- and post-assessment responses on Item 9 (Table 39) for individuals will illustrate the potential evolution of teachers' understanding. An examination of teachers' responses suggests that even though direct pre- to post comparison is problematic, several observations can be made: (1) teachers' responses varied widely; (2) some teachers' use of language evolved to include attention to formulating a researchable question or measurable conjecture; (3) attention to random assignment or random selection, sample size, and control groups was generally improved; (4) mention of conducting blind or double blind studies or designing to eliminate bias and lurking variables was new; (5) generally, connection to the context of the task was minimal.

The origins of these changes may be linked back to professional development Activities 1.2, 2.4, 2.5, and 4.2, for each of these attended to issues of experimental design, sample size, sampling bias, control, and randomness to some degree. Activity 2.5 using CPMP instructional materials introduced teachers to experimental design for purposes of establishing a linkage between cause and

Table 39

Item 9: Pre- and Post-Assessment Comparisons for Four Teachers

Code #		Level
20-pre	<i>Students should consider the environments of the people involved in the study. They would have to take people from many different countries and regions of our planet. They should consider the type of animals that each participant is around on a daily basis. They should also take into account any variables that might give false results. This type of thought is what I would be looking for.</i>	1
20-post	<i>Random; Large Sample; lurking variables not present; Blinded subjects & evaluators; <u>Hurrying!</u></i>	2
23-pre	<i>1) Random sampling; 2) Control/test groups; 3) Broad coverage; 4)</i>	1
23-post	<i>1) Clear concise question that is neither too broad, nor too specific (may first have to decide what "overweight" is. 2) Sufficient # of subjects in the study; 3) Random selection of people in the study; 4) Try to be either double-blind or at least single blind through this study</i>	3
50-pre	<i>--> Data they should consider collecting: *Name (I.D. purpose only), *age (see if there may be trends), *occupation (if any), *residence (city/farm), *Pets, *Known allergies, *Sex (Male/Female). Restating what the purpose of the study is.</i>	1
50-post	<i>1. Sufficient population, the study should have a sufficient # of data so that a non-biased view of the data can be obtained. 2. Randomness--subjects should not be picked by humans, rather computer generated listing. 3. Good definitions of what is to be measured and how it is to be measured (blind/double blind) so that bias can be taken out of the experiment</i>	3
17-pre	<i>*data collection method reasonable; *contains only 1 variable; *contains a control group; *unbiased; *accuracy of statistics; *appropriateness of method used to determine results; *conclusion & explanation.</i>	2
17-post	<i>Would need to consider: --how to randomly select sample to eliminate bias.--would need a control group. --large enough sample size.--Defined terms & question (what are they trying to determine).--appropriate statistics chose mean vs. median, st. dev. etc. --graph--appropriate to materials. --conclusions from graph.--randomization of mean diff or median diff. --p-value & conclusions drawn from it.</i>	4

effect. Many teachers later wrote reflections indicating they had never heard of lurking variables or blind or double-blind studies prior to Activity 2.5. The content of the CPMP investigations may have influenced teachers' view of overall

experimental design. Activity 4.2 may have been especially influential for reinforcing the connection between a big problem and a researchable question. During Activity 4.2, the Physician's Health Study, the big question of interest was "Do doctors discriminate against patients on the basis of weight?" The way in which the researchers operationalized the study in order to begin to shed some light on the question was by formulating a measurable question and designing a study that could logistically be conducted. A purposeful dialogue prior to investigating the data collection may have assisted teachers with this aspect of the statistical process.

Item 9's relative difficulty index increased from third most difficult to second most difficult from pre- to post-assessment even with change in group means of 1.66 to 2.32 (effect size 0.66). Gain-scores were significantly greater than 0. Teachers appeared to have increased their understanding of some basic concepts of experimental design, even though connecting concepts of experimental design to problem contexts remained somewhat elusive. Because of the very open-ended nature of this task, it may be that the task required more thought and time than that provided in a testing situation.

Analysis of Item 10

Item 10 corresponds to Tasks 14, 15, 16, 17, 18, 19, and 20 on the pre- and post-assessments. This collection of tasks comes from extensive research on undergraduate student understanding of sampling distributions (Chance et al., 2004).

This item posed a number of analytical difficulties. Initially, scoring was to be done by 7 correct→4; 5 or 6 correct→3; 3 or 4 correct→2; 1 or 2 correct→1; 0 correct→0. After carefully studying the multiple frameworks for this study and the work of Chance et al. (2004), the question was raised as to whether the number of tasks answered correctly could, by itself, determine the level of understanding of sampling distributions or the Central Limit Theorem (CLT). Additional analysis was undertaken using post-assessment scores to determine emerging patterns within the scoring categories. For example, did those scoring 5 or 6 correct do so in predictable ways? The short answer is no. Clearly to answer 5 or 6 correct, one must miss 1 or 2. There turned out to be several ways for that to happen in the data. The cases of 3 and 4 correct were investigated and found to have a similar lack of consistent pattern. To make matters worse, because the items were multiple-choice, it was impossible to know whether correct or incorrect responses were the result of guessing. Because the pool of seven questions is not quite representative of a binomial scenario (correct outcomes are not equally likely for each question—for some $p(\text{success}) = 0.33$, some 0.25, and some 0.20), a simulation using 10,000 trials was conducted using *Fathom2*. Assuming strictly guessing, the following probabilities were experimentally approximated:

P(7 correct) = 1/10,000	0.01%
P(6 correct) = 14/10,000	0.14%
P(5 correct) = 105/10,000	1.05%
P(4 correct) = 601/10,000	6.01%
P(3 correct) = 1876/10,000	18.76%
P(2 correct) = 3,207/10,000	32.07%
P(1 correct) = 2,995/10,000	29.95%
P(0 correct) = 1,201/10,000	12.01%

Thus, the likelihood that, by guessing alone, someone correctly responds to 4, 5, 6, or 7 questions is not very great. On the other hand, it would be quite likely that 1, 2, or 3 correct may be seen quite frequently if only guessing were involved.

In order to coordinate a four-level scoring rubric sensitive to the work of Chance et al. (2004) and others, the rules in Table 40 were used. Figure 30 displays the distribution of pre-assessment, post-assessment, and gain-scores for Item 10. It also includes a cross-tabulated display and summary of the ways in which teachers' scores changed from pre to post on this item.

Table 40

Item 10: Rubric for Scoring

Level	Rule
4	All 7 questions correct. It is very unlikely to achieve 7 correct answers by guessing.
3	Either 6 correct OR Correct answers for 14, 15, 16, (this suggests coordination for large sample size) and 20 (indicative of variability of sampling distribution decreasing with sample size) plus either a correct answer on #17 or a choice that suggests awareness of decreasing variability with sample size, plus either 18 or 19 correct. Other configurations of 5 correct were looked at on an individual basis to determine whether the combination of answers suggested a strong likelihood of understanding of sampling distributions
2	At least 4 correct, but not at the level required to score a 3. Deferring to the probability of 4 correct, it is unlikely that this would happen just by chance alone, but with 3 (or 4 in the case of 3 correct) incorrect, it becomes difficult to determine whether the incorrect answers are the result of guessing or of not knowing. 4 correct suggests at least a developing understanding of sampling distributions
1	Between 1 and 3 questions answered correctly. Guessing cannot be ruled out.
0	0 questions answered correctly

A positive shift is evident for this item. For example, 42 of 56 (75%) teachers scored at Level 0 or 1 on the pre-assessment; this dropped to just 17 of 56 (30%) on the post-assessment. This progress is especially promising, given that the professional development intervention did not directly address the essence of this set of tasks.

Furthermore, the literature suggests that understanding sampling distributions is particularly challenging (Chance et al., 2004).

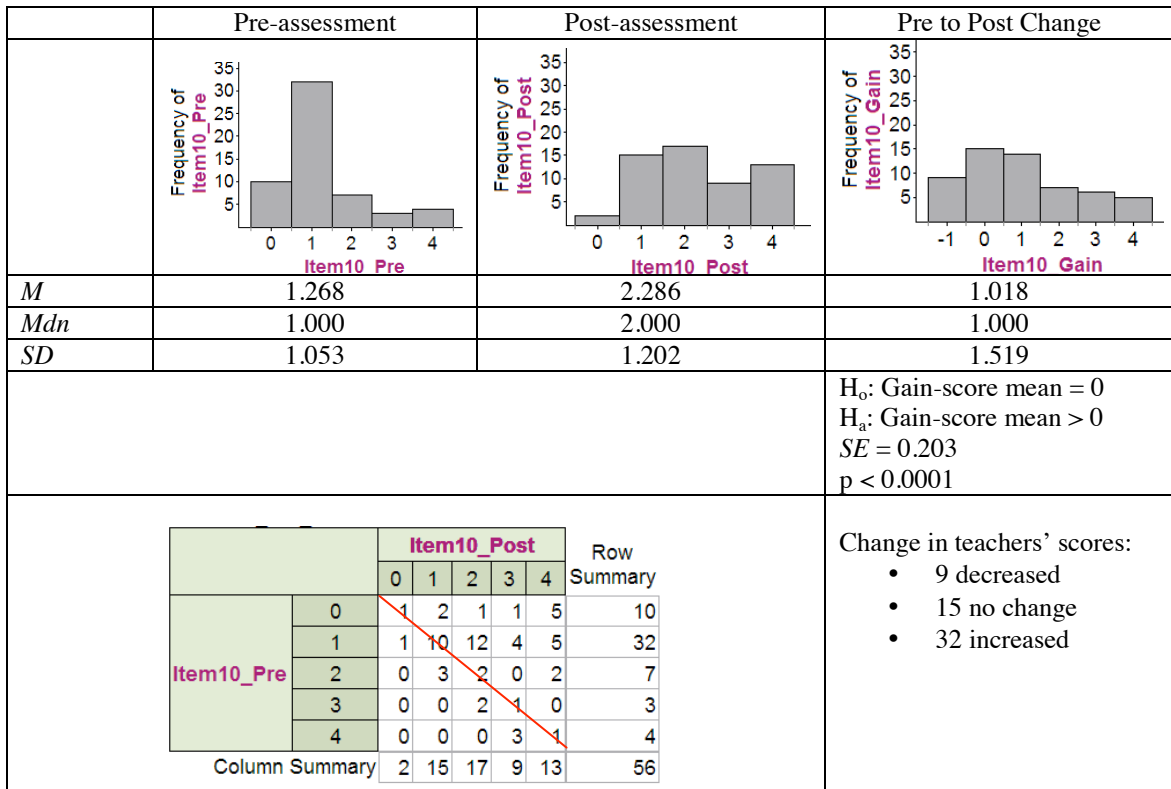


Figure 30. Item 10: Pre-, post-, and gain-score comparisons.

In order for a teacher to correctly answer all seven questions, s/he would have had to synthesize from one activity, Activity 3.4 on the last day of the session the essence of the Central Limit Theorem (CLT). The change in responses across this set of tasks suggests some type of powerful sense-making may have occurred. Because of the multiple-choice nature of the tasks, no rich written explanations were available to confirm or refute any conjectures. However, a task-by-task analysis is presented next in order to more carefully analyze the change across these tasks. Table 41 illustrates the change from pre- to post-assessment on each task. Tasks 18 and 19 were the least correctly answered tasks on the post-assessment. These two tasks, if answered correctly

would indicate an understanding that the sampling distribution of sample means becomes less variable and more normally distributed even for small sample sizes ($n = 4$).

Table 41

Item 10: Pre- and Post-Assessment Results from Tasks 14 through 20

	T14	T15	T16	T17	T18	T19	T20
Pre	22 (39%)	22 (39%)	21 (38%)	10 (18%)	19 (34%)	16 (29%)	28 (50%)
Post	42 (75%)	46 (82%)	35 (63%)	31 (55%)	23 (43%)	25 (45%)	47 (84%)

As Table 42 indicates, the number of teachers with 4 or more questions correct at the pre-assessment was 14 of 56 (25%), whereas at the post-assessment was 39 of 56 (70%). Because the questions were multiple-choice, random guessing was certainly a factor; however, the likelihood of someone scoring 4 or more by guessing is only in the neighborhood of 7%. Guessing would result in approximately four of 56 teachers scoring 4 or higher. Though it would be possible to estimate the probabilities associated with answering a certain number of questions correctly and guessing on the remainder of the questions, the evidence in this situation is strong enough to suggest that a positive shift in the distribution from pre- to post-assessment on these tasks occurred. Because of the complexity of the analysis of this item, and the many ways in which the data may be viewed, Table 43 contains seven two-way tables presenting the pre- and post-assessment results for each question.

Of the five teachers answering 6 correct on this set of tasks, the frequency of incorrect responses were T18: B—1, C—1; T19: B—1, C—2. Of the 10 teachers answering 5 correct on this set of tasks, the nature of incorrect responses were T16: B—2, C—1; T17: A—1, D—3; T18: B—4, C—2; T19: B—2, C—4; T20: D—1.

Table 42

Item 10: Total Number of Tasks Correct by Teacher

	0	1	2	3	4	5	6	7
Pre	10 (18%)	14 (25%)	5 (9%)	13 (23%)	5 (9%)	3 (5%)	2 (4%)	4 (7%)
Post	2 (4%)	2 (4%)	7 (13%)	6 (11%)	11 (20%)	10 (18%)	5 (9%)	13 (23%)
p(guess)	12%	30%	32%	19%	6%	1%	~0%	~0%

Table 43

Item 10: Pre- and Post Responses by Task (with correct answers in bold)

	14 Post					
14 Pre	A	B	C	D	E	Grand Total
A	11	1	1	0	1	14
B	3	1	0	3	0	7
C	4	1	0	0	0	5
D	2	0	0	0	0	2
E	16	3	1	1	1	22
blank	2	0	0	0	0	2
Don't know	4	0	0	0	0	4
Grand Total	42	6	2	4	2	56
	15 Post					
15 Pre	A	B	C			Grand Total
A	18	4	0			22
B	21	3	1			25
C	2	1	0			3
blank	2	0	0			2
Don't know	3	1	0			4
Grand Total	46	9	1			56
	16 Post					
16 Pre	A	B	C			Grand Total
A	13	7	1			21
B	11	3	5			19
C	4	0	3			7
D	2	0	0			2
blank	3	0	0			3
Don't know	3	0	1			4
Grand Total	36	10	10			56

Table 43—Continued

17 Post							
17 Pre	A	B	C	D	E	Blank	Grand Total
A	0	5	2	0	2	0	9
B	0	5	1	3	1	0	10
C	0	5	1	2	2	0	10
D	1	8	3	0	0	0	12
E	1	1	3	0	0	1	6
blank	0	5	0	0	1	0	6
Don't know	0	2	1	0	0	0	3
Grand Total	2	31	11	5	6	1	56
18 Post							
18 Pre	A	B	C	Grand Total			
A	7	10	2	19			
B	10	5	3	18			
C	3	5	4	12			
blank	1	3	0	4			
Don't know	2	1	0	3			
Grand Total	23	24	9	56			
19 Post							
19 Pre	A	B	C	D	Grand Total		
A	12	1	3	0	16		
B	1	3	2	1	7		
C	9	4	10	0	23		
D	1	1	0	0	2		
blank	1	1	2	0	4		
Don't know	2	1	1	0	4		
Grand Total	26	11	18	1	56		
20 Post							
20 Pre	A	B	C	D	Grand Total		
A	22	1	5	0	28		
B	6	1	0	1	8		
C	9	0	0	0	9		
D	2	0	1	0	3		
blank	4	0	0	0	4		
Don't know	4	0	0	0	4		
Grand Total	47	2	6	1	56		

Pairs of tasks missed by individual teachers included T17, 20: AD—1; T18, 19: BB—1, BC—1, CB—1; T17, 18: DB—2, DC—1; T16, 19: BC—2, CC—1. Of the 11 teachers answering 4 correct on this set of tasks, each had a different combination of 7 answers and not a single question was answered correctly by everyone. One interesting

signal in the data, however, was found in the number of correct answers per task with T14—8, T15—9, T16—7, T17—5, T18—1, T19—5, T20—9. Clearly Task 18 was the most missed task by this group. Task 18 was also one of two most missed tasks for teachers correctly answering 5 or 6 questions correctly as well. It was the task with the lowest percentage of correct responses overall. Task 18 asked teachers to respond to the following prompt:

Assume 500 samples of size 4 are randomly drawn from the POPULATION distribution shown on the previous page. I would expect the sampling distribution to be shaped more like _____.

Twenty-four of 56 (43%) teachers responded, “B. the population.” Though this response is indicative of a common misconception associated with sampling distributions of the mean, it is helpful to consider what may have happened during the professional development intervention which may have allowed some teachers to make the appropriate connection to the Central Limit Theorem, whereas other teachers did not.

Because Activity 3.4 was the single learning opportunity during the four-day session in which teachers purposely investigated the relationship between sample size, shape and standard error of the sampling distribution of the mean, it is possible that this incorrect response may have been an artifact of the classroom investigation, but for only some teachers. Specifically, during the investigation, teachers were instructed to create or find a population from which to sample (see Figure 8 in Chapter IV). This required them to locate an existing data set within *Fathom2*, go online and find an interesting data set to import into *Fathom2*, or create a population by defining an attribute using a formula (e.g., $\text{attribute} = \text{randomNormal}[54, 2]$). A number of groups in each of the three sessions opted to sample from a population that was approximately normally distributed to begin

with. Though a whole group discussion intended and appeared to develop a shared understanding that the pattern of behavior was common regardless of beginning population distribution, it may very well be the case that the interactional dialogue was insufficient to overcome individual experiences with sampling from normal distributions. For those teachers who began investigating sampling from normal distributions, the failure to experience the rapid convergence toward approximately normal shape for distributions of sample means with $n > 1$ may partially explain the answers for Tasks 18 and 19. Evidence for this theory was corroborated by a teacher during the post-interview (Site 3, Teacher David, Post-interview):

- SM: Okay. What . . . when you guys did this investigation on that last day, you had some population you were working with.*
- David: Hm, yeah.*
- SM: Do you remember that?*
- David: I think we generated actually a normal distribution and thinking back, we shouldn't have done that. It would've been more interesting to look at something that's skewed or very non-normal and, see . . .*
- SM: Oh, so because when you took your sample mean . . .*
- David: Yeah.*
- SM: . . . they automatically were . . .*
- David: They . . . they were all that the bell curve anyway, it's just it was being . . .*
- SM: Ah, okay.*
- David: . . . you know, compressed more, and more, and more, and more.*
- SM: Gotcha, all right. So you're thinking right now is that, if you take a sample of size four, it is not going to get . . . it is not going to get normal very fast enough because it's so weird.*
- David: Yeah, that's . . . that's what I was thinking the last day, but now, when I look at this and that, maybe it's . . . it . . . it looks very close to what it was originally. So I'm thinking maybe I should have chosen something different.*

One thing that was particularly thought-provoking about this interview response was that this teacher teaches statistics to high school students. As evidenced several times during the interviews and professional development, though his knowledge of statistical content

appeared stronger than his peers in the session, his understanding of the content appeared more procedural than conceptual and this procedural understanding may have been somewhat resistant to change.

It is conjectured that teachers able to answer all seven questions correctly may have been those who began the sampling activity with non-normal distributions to begin with. This suggests a possible design change for further work in this area. Future research may intentionally investigate the differences in performance on these items depending on what types of initial populations were sampled. Perhaps the stipulation that all groups investigate different non-normal distributions during this activity may provoke the desired disequilibrium for more of the participants. Alternately, an extension to the activity in which teachers are asked to carefully reflect on the experience and formalize some of their thinking or apply it in another situation might be beneficial to support teachers' understanding.

Given that sampling distributions and the CLT are notoriously challenging for students to understand, only a small number of activities in the professional development directly targeted sampling distributions, and that teachers in this study made significant gains toward understanding sampling distributions as evidenced by the written assessments, the investigations underpinning this progress show promise for supporting teacher learning. Figure 31 displays the comparison of the number of questions correctly answered on the post-assessment (horizontal axis) to teachers' scores on the pre-assessment (vertical axis). The circled values represent teachers who scored 0, 1, or 2 on the pre-assessment and then answered four or more questions correctly on the post-assessment.

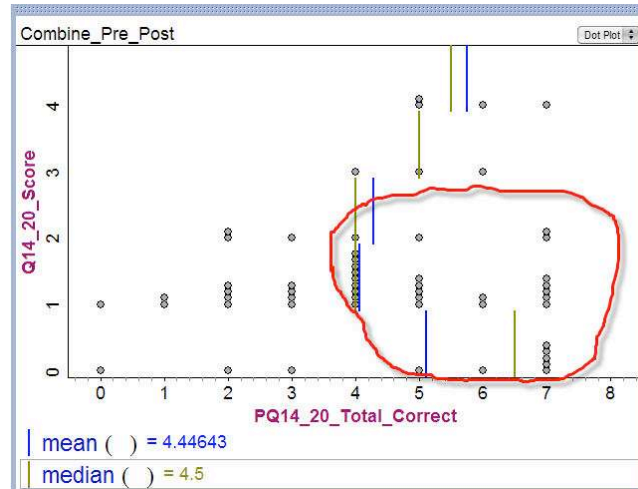


Figure 31. Item 10: Comparison between number correct on post-assessment (horizontal axis) to pre-assessment score (vertical axis).

Item 10 was judged the most difficult item for teachers on both assessments. Nonetheless, scores grew from 1.27 to 2.29 (effect size 0.67) and gains scores were significantly greater than 0. When post-assessment scores were disaggregated across prior statistics background, Figure 32 suggests that prior statistics background was not a good predictor of performance on this item on the post-assessment.

Summary of Item-Specific Analyses

Ten items designed to assess teachers' understanding of comparing distributions were administered before and after a four-day professional development intervention. Comparing distributions was conceptualized to incorporate big ideas of distribution (Items 1, 4, 5, and 7), variability (Items 2, 3, and 8), and sampling distributions (Items 6, 9, and 10). Four items were included for their potential to contribute to the assessment of teachers' understanding of distribution, three items for variability, and three items for sampling distribution. For all ten items, mean teachers' responses indicated significant improvement in understanding from pre- to post-assessment. Table 44 displays the results

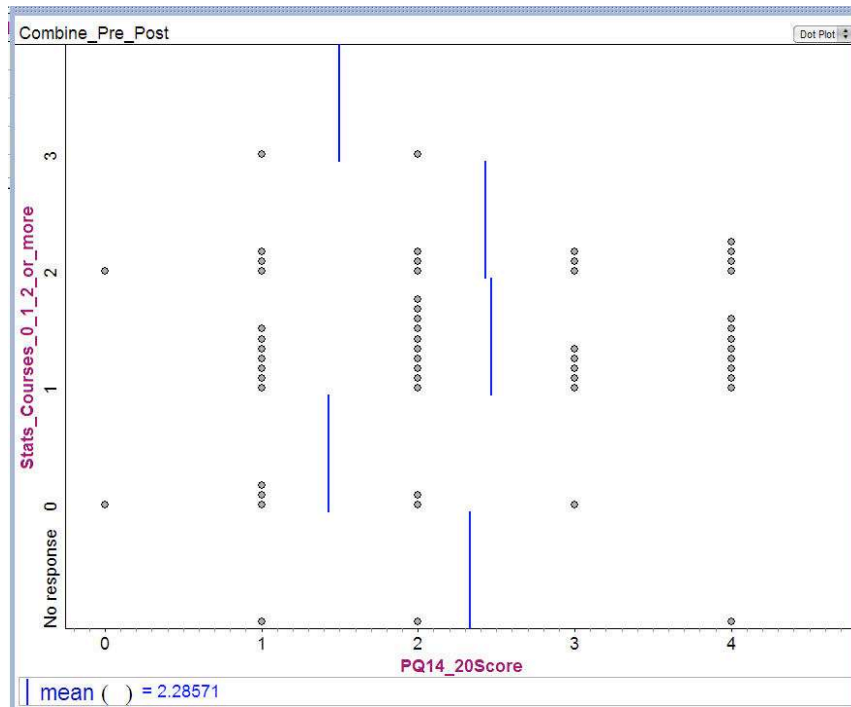


Figure 32. Item 10: Post-assessment scores disaggregated by number of statistics courses taken.

of the assessments when items are aggregated by big idea. This tabular representation suggests that teachers' understanding of distribution and variability were similar at the beginning and end of the intervention, each with similar gains. The concept of sampling distribution appeared more challenging for teachers at the pre-assessment and the improvement on this concept was greater than the other two major concepts. Matched pairs gains-scores across all three big ideas were significantly greater than 0.

As part of the assessment of teachers' content knowledge, a brief survey of statistical comfort-level was conducted. Table 45 presents teachers' self-reported comfort-level for each of seven major statistical contexts before and after the professional development intervention. Each indicator was rated using a 5-point scale with 1 being weak, 5 strong. Pre-post survey responses were matched by participant and gain scores computed. In each category, gain scores were significantly greater than 0 with

Table 44

Pre-, Post-, and Gain-Score Comparisons by Big Idea Cluster

		Distribution 4 Items	Variability 3 Items	Sampling Distributions 3 Items	Total 10 Items
Pre	<i>M</i>	2.05	2.01	1.57	1.89
	<i>Mdn</i>	1.94	2.00	1.50	1.85
	<i>SD</i>	0.56	0.70	0.66	0.48
Post	<i>M</i>	2.89	2.84	2.58	2.78
	<i>Mdn</i>	2.88	3.00	2.67	2.88
	<i>SD</i>	0.51	0.73	0.71	0.54
Gain	<i>M</i>	0.83	0.83	1.01	0.89
	<i>Mdn</i>	0.75	0.67	1.00	0.88
	<i>SD</i>	0.54	0.84	0.80	0.49
	t_{55}	11.64	7.459	9.419	13.69
	<i>p</i>	< 0.0001	< 0.0001	< 0.0001	< 0.0001
	Effect size	1.54	0.99	1.26	1.82

$p < 0.0001$ using Wilcoxon signed rank tests. These results suggest that participants' personal beliefs about their statistical understanding significantly increased as a result of the professional development intervention. The asterisk (*) represents a gain score significantly greater than 0. Note: Several ratings were recorded in $\frac{1}{2}$ point intervals, such as 1.5, 2.5, etc. For presentation in the table below, pretest scores were rounded up to the next highest integer and posttest scores were rounded down. These adjustments affect mean score calculations slightly.

Summary of Content Assessment Results

Analysis of content assessment data suggests that the three groups of teachers involved in this study began with a similar understanding of comparing distributions and ended with a stronger, but comparable understanding of distributions, regardless of professional development site attended. Statistically significant gains were seen from pre- to post-assessment for each item on the assessment, for overall scores on the assessment,

Table 45

Teachers' Self-Reported Comfort Level with Statistical Ideas at Pre- and Post-Assessment

	Pre-assessment (n=56)						Post-assessment (n=56)						GAIN
	1	2	3	4	5	M	1	2	3	4	5	M	
Descriptive statistics (mean, standard deviation, z-score)	1	17	19	14	5	3.09	0	2	15	25	14	3.91	0.82* (n=56)
Statistical Graphs (histogram, boxplot, bar graph)	0	3	11	23	19	4.04	0	0	5	22	29	4.43	0.39* (n=56)
Distributions (normal, chi-square, probability density functions)	28	15	10	2	1	1.80	4	16	25	9	2	2.81	1.01* (n=56)
Experimental Design (surveys, blocking, bias, sampling methods)	21	21	9	4	1	1.98	3	8	10	26	9	3.54	1.55* (n=56)
Correlation and Regression (least squares, r^2 , residuals, outliers)	21	22	10	1	2	1.95	7	14	20	9	6	2.88	0.94* (n=56)
Sampling Distributions (Central Limit Theorem)	42	11	1	0	2	1.37	9	12	21	7	7	2.86	1.49* (n=56)
Statistical Inference (t-tests, confidence intervals, chi-square tests, power, Type II error, ANOVA)	45	6	2	2	0	1.29 (1 blank, n=55)	26	20	8	1	1	1.79	0.47* (n=55)

* $p < 0.0001$.

and for big idea clusters on the assessment. Growth in understanding, as measured by the assessment instruments, suggests that the professional development experience described in this study was broadly accessible and applicable, regardless of prior statistics education. Teachers' self-reported comfort-level with statistical ideas improved pre- to post, suggesting consistency between beliefs about statistical understanding and performance on statistical comparing distributions tasks.

Interview Results

As described in Chapter III, three teachers from each of the three sites were interviewed prior to the professional development. The average number of courses taken and their average comfort-level, as reported on the initial pre-assessment, were remarkably similar to those representing all 56 teachers in this study. Five of 9 of the teachers reported never teaching statistics, while 3 teachers reported teaching a statistics course and 1 teacher reportedly taught some statistics during an algebra course. The gender make-up of the interview pool is more heavily weighted toward females largely due to two factors: (1) more than half of the teachers in the study were women (31/56, 55%), and (2) no men from one of the professional development sites were willing and able to participate in both pre- and post-interviews. Table 46 summarizes some of the characteristics of the interviewed teachers.

Table 46

Characteristics of Interview Teachers

Name	Site	Gender	# years teaching math	# stats courses taken	Aggregate comfort-level	Ever taught statistics?
Cameron	1	M	16	1	1.71	N
Jessy	1	F	8	1	2.29	N
Lorraine	1	F	2	1	1.57	N
Jaylee	2	F	1	2	1.42	N
Sasha	2	F	5	1	2.14	Y
Jordan	2	F	3	1	2.83	Y
June	3	F	10	1	1.57	In algebra
Callie	3	F	18	1	1.57	N
David	3	M	8	2	4.57	Y
Average, $n_{\text{interview}} = 9$			7.89	1.22	2.19	
Average, $n_{\text{all}} = 56$			10.13	1.18	2.22	

Each teacher participated in clinical interviews approximately one month before and eight days following the professional development program. Interviews ranged from

about 30 minutes to 84 minutes, with a typical time of 45 minutes to one hour. All interviews were transcribed for analysis. Transcripts totaled approximately 800 double-spaced typed pages which were then coded in chunks representing entire responses to questions. Further coding associated with the framework coordinating *comparing distributions*, distribution, variability, and sampling distributions and their various dimensions. As discussed in Chapter III, this procedure represented a compromise between an entirely *a priori* coding approach and a completely inductive approach. Throughout the coding process, codes were created or amended as needed in order to give voice to the data. Data and analysis from the pre- and post-interviews will be presented next.

Descriptions of Interview Teachers and Pre-Interview Summaries

This section contains descriptions of each of the nine teachers interviewed for this study along with summaries from the analysis of the pre-interviews. The interview prompts may be found in Appendix C. The section is organized according to the site of the professional development; each site contains summaries of three teacher interviews. The sections that follow serve to provide a richer sense of the nature of the background and statistical content understanding articulated by the teachers in this study. Furthermore, these data in conjunction with the previously reported pre-assessment data served to inform the development of the HLT that was described in Chapter IV. The section will conclude with a brief summary of the interviews.

Site 1 Teacher Pre-Interviews

Cameron. The pre-interview with Cameron lasted approximately 53 minutes.

Cameron is an experienced mathematics teacher with more than 16 years of teaching in a small rural school district. The description of his formal statistical experience in college reflected frustration and difficulty. One of the first things he said during the pre-interview was, “I’m terrible at statistics” (PreI, Cameron). He had been advised to take a statistics course for which he lacked the prerequisite Calculus II course. He described struggling with integration by parts in order to derive various probability distributions, while working very hard at understanding the concepts. He earned a C in the course and characterized the grade as a gift.

And so when, when it was all said and done, I felt that I had learned very little. I, I was just so overwhelmed that what I took away from the course was, there wasn’t a whole lot of really, content that I learned, and consequently, I, I just had a phobia in statistics. I never took another course. I probably learned z-scores and normal distribution, but I remember poisson distribution, chi-square distribution. I have since lost it and, and I always was very, I, I was doubtful as to how and when to apply those distributions. I mean, I, I struggled so much with, with deriving the formulas for them, that I was lost in the woods and I couldn’t see the big picture. Even to this day. (PreI, Cameron)

Cameron teaches the “very fundamental basics” (PreI, Cameron) of mean, median, mode, boxplots, z-scores, and normal distribution in Algebra I and II.

When comparing distributions, Cameron’s responses indicated his attention to measures of center as well as spread.

Well, I would look for grouping first and the . . . the closeness of the grouping, the dispersion like this one is more widely dispersed than the other one. This is much more closely. This . . . this I think is a greater central tendency here that . . . whereas this one has a wider dispersion of data. (PreI, Cameron)

He referred to outliers as “glips” and referenced “glips on a radar screen” (PreI, Cameron). He preferred to reason from dotplots because the actual data were preserved for his view. His second graphical preference for comparing distributions was histograms and he least preferred boxplots. He acknowledged that he had only learned about boxplots during the last 10 years and did not have great experience interpreting distributions with the boxplot representation.

When asked about variability, Cameron used the language of “diversity” and “how far it is away from . . . from the central tendency, from the . . . from the . . . the middle of the road” (PreI, Cameron). His reasoning and statements indicated a standard deviation perspective. He did not remember the formula for standard deviation, but appeared to know that standard deviation was sensitive to sample size. His responses to pre-assessment Tasks 10-12 indicated that his reasoning about variability included a view of “conformity,” indicating less variability and attention to range as a measure of spread. His written responses lacked sufficient clarity to score highly. He did correctly identify all distributions with greater variability but his written reasoning did not reflect a clear statistical perspective.

Cameron’s view of significant difference was largely context-bound and he suggested that comparing groups of unequal size was problematic. He appeared to understand that smaller samples tend to have larger variability. When asked about the experimental design task, Cameron wanted a control and treatment group, but did not attend to the formulation of a researchable question. He was able to interpret the performance task simulation results as intended.

Though Cameron lacked confidence in his understanding of statistics generally, his responses were indicative of a statistical perspective. He attended to measures of center while coordinating standard deviation. His view of comparing distributions appeared to coordinate center and spread but may have been limited to a view of distributions that were normally-distributed to begin with. His overall pre-assessment score of 1.85, with demonstrated level of understanding of 1.88 for distribution, 2.33 for variability, and 1.33 for sampling distributions reflects his relative strength of reasoning using standard deviation as a measure of variability while attending to “central tendency” when comparing distributions. His responses to the written assessment may reflect his lack of confidence in his statistical reasoning ability.

Jessy. The pre-interview with Jessy lasted approximately 53 minutes. Jessy had been teaching high school algebra and geometry for eight years at a relatively large rural school. She described her undergraduate statistics course experience as

It was taught by a professor who we did not . . . I think we did have a textbook, but he handed us this packet, at the beginning of the class, and it was all his notes which would like go off the page and then be like, “See another page and . . . ” It was just crazy, I guess. I mean, it was not . . . I didn’t care for his teaching style I guess. And he wasn’t organized and it just . . . afternoon class and I did okay in the class. I mean, I think I probably got an A because I didn’t . . . there weren’t too many classes that I didn’t get As in, but I felt . . . I didn’t feel like . . . it was one of those classes where I got the A, but I didn’t feel like I earned the A, you know, I didn’t feel like my understanding was what it should have been. (PreI, Jessy)

When asked about things she learned really well, she responded:

Probably like, you know . . . you know the curve and then, probability, the area to the left or the area to the right, that type of thing. Calculating means, standard deviation, but all those different . . . I don’t know . . . remember the type of test, but all the different types of test that you do . . .

I don't remember. Yeah, I just have a basic, basic concept of stats. (PreI, Jessy)

When asked about things she may have had difficulty with, she responded:

I learned it well enough to do okay on the test, but not that I ever had the big picture, you know, all the . . . together. And I guess I had difficulty knowing when to use what test for what type of . . . and I don't think he ever gave us data and said, "You decide what test to do." Like, the test that we took was very, "Use this test to determine, you know, so forth." So it was not really . . . it was more . . . I don't know how to explain it. He . . . he gave you hints on where you were going and so therefore, I knew how to do that procedure, so I did it and I did well, but I don't think I ever understood . . . (PreI, Jessy)

At a later point in the interview she discussed her experience in a graduate education course in which she conducted a small study in her own classroom. She explained the meaning-making that was important to her from the experience:

That's probably about the only one that I've had doing my own collection of data. That's probably what I remember more from than anything that I ever learned in college because I think in college it was like, "Here's a story problem. Here's a story problem. Here's a story problem." And I memorized a method on how to do it. I don't understand why the method . . . I did the method, but at that point, I was just like, "I just want to get through the class," you know . . . and . . . and so it didn't have a lot of meaning for me, but when I did this research study for a class in my graduate studies, that . . . had a little bit more meaning, you know. (PreI, Jessy)

When comparing distributions represented as histograms, Jessy attended to shape, center, and spread. She compared histograms using the metaphor of looking for "high mountainous, small range . . . with resemblance to a bell curve" (PreI, Jessy). When comparing distributions represented as boxplots, Jessy focused on the interquartile range and the whiskers to determine the presence of outliers. She appeared to understand that if the boxes were similarly located, then there may not be a significant difference in the means of the distributions, but she was not confident in her reasoning. As she compared

distributions as dotplots, her reasoning seemed to reflect a counting of dots approach and reasoning from individual data values as opposed to distribution. She was concerned with statistical significance; however, she was unable to make the connection as she had with the boxplot example. She did not attempt to apply the boxplot approach to the problem. She acknowledged her preference of representations from which to reason was first histograms, then dotplots, and finally boxplots. Jessy discussed little familiarity with boxplots from her undergraduate coursework; she graduated in 1997.

Significant difference to Jessy seemed to depend upon sample size and she thought samples of size 21 and 23 were too small to confidently conclude anything. Her reasoning appeared inconsistent across representations. When reasoning about variability, she indicated sensitivity to both the range of the data and the consistency of frequencies in a histogram. She displayed a “variability as bumpiness” approach and also attention to sample size and concentration of data. On the experimental design task (Item 9), Jessy required additional prompting and support to engage with the question. Her written response was generic and did not deal with the context provided in the task. She seemed to understand that attention to sample size and random “populations” were important and her verbal response suggested that a graduate course experience may have been informing her thinking about this. The formulation of a researchable question was not apparent in her written or verbal responses. On the performance task, Jessy was not able to reason beyond the graphical shape and center of the distribution to respond to the question asked. With some prompting, she was able to ultimately make necessary connections to respond to the task; however, her initial response suggested that reasoning from this type of simulation data may have been foreign to her.

Jessy's overall pre-assessment score of 1.85, with demonstrated level of understanding of 1.38 for distribution, 3.00 for variability, and 1.33 for sampling distributions may have under-represented her understanding of these statistical ideas. During the interview, it appeared Jessy was concerned with trying to generate the "correct answer" and her confidence was not high. She indicated an attempt to determine what the researcher was looking for and did not appear sure of her responses.

Lorraine. The pre-interview with Lorraine lasted approximately 35 minutes. Lorraine had been teaching at-risk students in a relatively large rural school for two years. Lorraine described her undergraduate statistics experience from eight years ago as

I remember it was really fast and furious, you know, it was just thrown as a semester course that was just thrown at you. And it was pretty easy math, but there was just so much to remember, you know, I think probability and statistics is pretty easy, like the actual crunching of numbers, is pretty easy, but there's, it's just, it was just so much, you know, with all the different tests, and when you use this, and how you do this, and all this . . . (PreI, Lorraine)

When asked whether there were things she remembered well, her reply revealed both familiarity with statistical language as well as confusion:

No, mean, median, mode, you know, simple. I remember like t-squared, z-squared, you know, stuff, I don't, chi-squared. Don't ask me when I would use them, you know, I know I could pick up a book and be right back into it, yeah, but, I don't know. I can't say there's a heck of a lot I remember from it, and I know that's awful. (PreI, Lorraine)

When comparing distributions represented as histograms, Lorraine mentioned shape, center, and spread. Her understanding of standard deviation appeared to be heavily influenced by the range of a distribution. Her idea of shape seemed to hinge on normal distributions that she referred to as "bell-curves." When a distribution was not "a perfect bell," she called it "skewed." When reasoning from boxplots, Lorraine mentioned her

lack of familiarity with the representation and as she referred to the mean as Q2 while pointing to the median, calling it the middle line. She acknowledged that boxplots were her least favorite representation due to her lack of exposure to them. Her favorite representation was the dotplot because of the preservation of the actual data values and she declared, “I like my numbers.” The context of the dotplot comparison task appeared to heavily influence Lorraine’s comparison. She referred repeatedly to her experience as a teacher when conveying her response.

With respect to her understanding of variability, Lorraine’s reasoning seemed to be idiosyncratic. On pre-assessment Tasks 10-12 (Item 8), although she correctly identified the distribution, from the pair in each task, with more variability, her reasons incorporated variation in frequencies, bumpiness, and range and she appeared to equate those showing less variability with normal distributions. Lorraine’s idea about significant difference appeared heavily dependent upon the context and the magnitude of the difference; she did not demonstrate a statistical view of significant difference. Her responses focused exclusively on measures of center and context. When responding to the question about experimental design and a high quality student response, Lorraine, like many others, provided a generic response. She realized that the number of subjects and concise definitions for variables of interest were important. She seemed to believe that questions with yes or no answers were preferable. She indicated that her students would likely struggle with a task like this. On the performance task, it appeared that Lorraine was unfamiliar with reasoning from simulation data. She was unable to connect the situation to the simulation without prompting.

Lorraine seemed to have some ideas about statistical measures and making comparisons. Her understanding appeared to be heavily based upon context and experience, but she was also aware of some statistical conventions, such as looking at shape, center, and spread. It did not appear that her conceptions were well-connected. Her overall pre-assessment score of 2.00, with demonstrated level of understanding of 2.00 for distribution, 2.33 for variability, and 1.67 for sampling distributions was consistent with her interview responses.

Site 2 Teacher Pre-Interviews

Jaylee. The pre-interview with Jaylee lasted approximately 36 minutes. Jaylee seemed reserved and hesitant or nervous during the interview. She had just finished her first year teaching high school mathematics, although she had been teaching for seven years. When describing her statistics course experience, she mentioned taking one undergraduate and one graduate course. The graduate course included using the computer but she could not recall in what ways; she dropped the course because of time restrictions. She reflected on her experience as

Mostly it was just a textbook and we just grunted it out through the chapters . . . I guess just calculating things comes pretty easy to me, so as long as it's presented to me at the time I can do it. I can't replicate it later, because I just don't remember. (PreI, Jaylee)

She thought high school students should know how to read and interpret graphs and be critical of what was being represented in order to not be misled. She included mean, median, and mode as important and “knowing which test is better for which circumstances” (PreI, Jaylee).

When comparing distributions using the histogram representation, Jaylee focused on the range and the mode. When comparing using boxplots, she coordinated the use of the median and the mean in the context of the problem to make comparisons but did not appear to take variability into account. When comparing with dotplots, she located the median and the mode for both distributions. She mentioned that the sample sizes were close, but had reservations about the reliability of the measure for making decisions. When asked about her preference of representations, she preferred boxplots, then histograms (which she referred to as bar graphs), and then finally dotplots, but she did not articulate a reason for her preferences, even when prompted.

When asked about variability, she said “variability means they’re not all the same.” On the pre-assessment, Jaylee left Tasks 10-12 (Item 8) blank. When asked about the tasks during the interview, she communicated a view of variability as “bumpiness.” Statistically different seemed to be associated with standard deviation to Jaylee as she said “The more standard deviations you get away from the median, the less likely it is the norm for the group” (PreI, Jaylee). Furthermore, the magnitude of differences in centers from distributions seemed to influence her view of significant difference.

On the pre-assessment Task 13 (Item 9) about the high quality student response to an experimental design task, Jaylee responded, “I’m not sure” (Pre, Jaylee). During the interview, she talked around the task, but never provided a response that could be considered on-target for responding to the task. She did mention a study she had heard about that was related to the context of the problem, but she was unable to articulate her own ideas for the response. On the performance task Jaylee was unable to connect the situation to the simulated data in order to respond to the task without significant

prompting. She mentioned that her work with simulations was limited to geometric probability, but she remarked that all of that was being removed from the curriculum in her school in order to make room for algebra.

Jaylee received the lowest score of all 56 teachers on the pre-assessment. Her overall pre-assessment score of 1.00, with demonstrated level of understanding of 2.25 for distribution, 0.33 for variability, and 0.00 for sampling distributions suggested a very basic understanding of descriptive statistics. During the interview, she revealed some knowledge of variability as range and standard deviation, although she connected standard deviation to comparing distributions only through the language of significant difference and the connection was problematic. Her lack of confidence was evident throughout the interview.

Sasha. The pre-interview with Sasha lasted approximately 47 minutes. Sasha had been teaching mathematics in a mid-sized rural school for five years. She was the only teacher who described her undergraduate statistics experience as including student activities. Specifically, she remembered sampling from a bag of candy, collecting data on number of steps needed to traverse a hallway, comparing heights and gender, and making different distributions. Use of technology was not part of the course and she said, “It was, you know, pencil and paper manipulation and, you know, doing different distributions and different types of plots, and that kind of stuff” (PreI, Sasha). She remembered the Central Limit Theorem as something that she never fully understood.

She compared histograms by focusing on the bars with the highest frequency as a measure of center and used language of skewness and bell-like to describe distribution shape. She also compared ranges. She compared boxplots by comparing quartiles, length

of whiskers, and looked for outliers. She alluded to a balancing approach where an outlier might be countered by another data value. When comparing dotplots, her focus began with range and went to some sort of center peak. The context of the dotplot problem was clearly able to be connected to her teaching practice and her experience contributed to her analysis and conclusion.

Sasha was initially unable to articulate a meaning for variability in the context of data and statistics, and acknowledged that pre-assessment Tasks 10-12 (Item 8) were very difficult for her. After referring to her answers from the pre-assessment, she offered a model for her thinking about variability as, “I would say how much the data differs from what it’s with.” When asked to apply that model to a situation, she said,

Because of the way this peaks, it’s kind of like you know the majority of the people are here [pointing to T10B]. We’ve got some over here, and we’ve got some over here. Where here [T11B], with it going up and down, you can’t really say well this is where, you know, the majority of our people are. You know, this person is different than, you know, these people. There’s a greater difference between these two people than there is between these two people. And it kind of, where this builds, and then levels, this kind of jumps around. (PreI, Sasha)

When trying to apply her definition to Task 12, Sasha said,

Looking at this one [T12B], this one jumps around a lot more, like this one does [T11B]. But, this [T12A] is more spread out than what this [T12B] is. This one’s [T12B] kind of a little more clustered together. So I think that is why it was so hard for me to kind of look at these two and determine which has more variability. (PreI, Sasha)

On the pre-assessment, Sasha correctly selected distributions with the most variability in two of three situations, Task 11 and Task 12; however, her explanations suggested that she was reasoning from a perspective where she coordinated “bumpiness,” “frequency,” and “range” to determine relative variability. She was clearly not viewing variability as the average distance from a central anchor point, as her Task 10 incorrect choice of

“Group 2 because Group 1 has the same number of people in each category for number of quarters” (Pre, Sasha) represents.

Additional interview responses suggested her view of significant difference had mostly to do with the magnitude of numerical differences of measures of center and depended heavily on context. When asked about characteristics of a carefully designed study, Sasha was hesitant to respond and suggested she had very limited knowledge about designing studies. Finally, on the performance task incorporating results from a simulation, Sasha did not connect the simulated results to the context of the task and had difficulty determining how to proceed.

Generally, Sasha’s interview responses were suggestive of a teacher with some prior statistical knowledge, and her understanding of statistical ideas appeared to include basic ability to read and interpret graphs, basic understanding of measures of center and spread, and reasoning from the context of a problem. Her understanding of comparing distributions seemed heavily dependent upon context and though a distributional view of data was communicated, she did not communicate a well-coordinated statistical perspective. Her overall pre-assessment score of 1.70, with demonstrated level of understanding of 1.75 for distribution, 2.33 for variability, and 1.00 for sampling distributions was consistent with her interview responses.

Jordan. The pre-interview with Jordan lasted approximately 39 minutes. Jordan had just completed her third year of teaching high school mathematics. She had taken one undergraduate statistics course in college. She mentioned that her statistics professor was foreign and consequently “she didn’t understand a thing” (PreI, Jordan). Additionally, she described the experience as

It was lecture format. Come in, 40 minutes of a lecture, 40 minutes to an hour, go home, do your homework, come back. Take a quiz. No technology integrated. It was like board and paperwork the whole time. No excel spreadsheets or anything like that. Mostly theory. (PreI, Jordan)

She could recall studying populations and samples and chi-square tests, although she said she never understood chi-square tests, but she was uncertain as to whether she could attribute her statistical knowledge to the college course or from teaching herself when teaching *Functions, Statistics, and Trigonometry* (FST) in high school. She acknowledged that she teaches standard deviation, correlation, and least squares regression, but relies on technology for computation and does not teach the theory behind any of the concepts. She feels that all students should know how to read and critically interpret graphs, but suggested that perhaps not all students should need to know how to make graphical representations; to Jordan, it was more important that students could read and interpret graphs rather than construct them. She spends time with students understanding differences between median and mean for summarizing data in groups.

When comparing distributions as histograms, Jordan looks for clumps or groups of data. She locates the medians. She mentioned considering the minimum and maximum values, and whether the distributions were skewed, although she was uncertain of the direction of the skewing. When comparing with boxplots, she immediately looked at the median and the corresponding quartiles between distributions. She appeared to associate clustering of data with shorter quartiles and her reasoning appeared well-developed.

When comparing distributions with dotplots she immediately located the median of the control group and used it as an anchor for comparison with the treatment group. She compared the number of students scoring above 42 in both groups and the number of students scoring below 42 in both groups. She disregarded a value she considered to be

an outlier and then referred to the context of the problem to provide a reasonable conclusion as well as several reservations for her conclusions.

Jordan explained variability as “how spread out the data is, versus grouped together in one area” (PreI, Jordan) and referred to range and IQR. She said she did not like pre-assessment Tasks 10-12 (Item 8). Of the three tasks, Jordan correctly identified only one distribution with more variability than its comparison distribution (Task 11). All three of her reasons seemed to indicate she was looking for the proportion of the data above and below the median. Symmetric distributions appeared to imply the same variability to Jordan. During the interview when referring to Task 12, she says,

So this, what I had difficulty with, this, these two are different in a way that these are very symmetric [graph A], these ones [graph B] you have this jumping pattern, which maybe isn't variability. But is, I don't know, there's kind of gaps. There's something there, you know you don't have this nice symmetrical graph, versus this one that's choppy. So, but I don't know if that's variability or it's another term.

Later in the interview while referring to pre-assessment Task 12, she selects School B as the distribution with more variability, and says,

Data points are farther from the median. So, I looked at the median. And then I compared how far the data points were from the median. And these ones had a tendency to be farther away from the median than these ones on average. Yeah, so if the median here is in the middle then those are relatively more close to the median than this one. So it's like I put a stake at the median and see how far each point is away. [Why the median?] Umm, because that's the middle data point. And I never use mean when comparing those things because if you have one value that's really high or really low, it pulls the mean out of the middle.

Jordan appears to understand the importance of a central anchor point, but has not coordinated standard deviation into her thinking about variability, even though she teaches standard deviation to her students. She may be thinking of mean absolute deviation from the median, but that is unclear from the interview.

Jordan's view of statistically significant difference appeared largely dependent upon context, but her reasoning seemed to suggest an understanding that if data were piled up near the median of one group and similarly for another group, if the medians were different, then there probably was a significant difference. When presented with examples from which to reason, her choices were inconclusive; she waffled between using context to determine significance and then the magnitude of the difference. She did relate significant difference to a change that was relatively large compared to the scale, for example,

You know if the AP calculus exam, and everybody moves from a 3 to a 3.2 average on a five-point scale, then that might be significant, where a five percent in the class is less significant. (PreI, Jordan)

When responding to the experimental design task (Item 9), Jordan expressed the need to clearly define and identify variables, the need to have variables that were measureable, and the need for large random samples. She did not volunteer that the formulation of a researchable question was necessary and when prompted, she suggested that the question in the task was researchable. She mentioned that students in her school do not have experiences conducting this kind of research, neither in mathematics nor science classes. On the performance task, Jordan was able to interpret the problem appropriately, but needed some prompting in order to connect the simulated data with the context of the problem. She acknowledged that simulations were only a very small part of her curriculum.

Jordan's interview suggested her experience teaching FST may have supported her statistical understanding. She communicated a generally sound understanding of measures of center and attention to shape and spread. Her concept of standard deviation

was problematic and her experience applying statistical reasoning in task-based situations was limited. It is possible that her reliance on a textbook limited her vision of statistics. Her overall pre-assessment score of 2.20, with demonstrated level of understanding of 2.75 for distribution, 2.33 for variability, and 1.33 for sampling distributions was consistent with her interview responses.

Site 3 Teacher Pre-Interviews

June. The pre-interview with June lasted approximately 39 minutes. June had been teaching grades 7-12 mathematics for 10 years in a very small rural school district and had taken one statistics course in college. The description of her statistics course experience reflected challenges she had faced while learning statistics.

Well, to be honest with you, I had the class twice, because the first time I took it I had never had any stats in my life. You know, I didn't even know what $\bar{x}$ was. And I was actually doing very well in the class, you know getting through the formulas but I just didn't have a clue. I didn't understand because the teacher just, you know, threw up all the symbols, and constantly kept talking about $\bar{x}$, and I actually withdrew from the class, because I just felt lost. Even though, actually, you know academically the grade was an A. But, so then I re-took it with a different teacher and it was a man who explained, you know, used more words. In fact in his handouts it was all written out, you know, what all these symbols stood for and in the context of the problem. You know, so I really liked that a lot better, you know all the symbols made sense. And so I remember that specifically that there was a lot of language, the English language that went with it. [Were there any just concepts you felt like you really struggled with or that you never really grasped?] Well I guess the whole $\bar{x}$, the mean. Yeah, I guess that's what I was missing the first time around. And I think maybe the teacher assumed that everyone even had some background, even though there was no prerequisite for the class. So maybe she was teaching, you know, as if everyone understood her . . . [about the second time in the course] Well as I was taking the class, you know I did very well in the class, and I understood everything, but I just haven't really had to teach stats since . . . I remember t -scores, and z -scores, and all that stuff, it's just that I don't really fully understand what it means. And even this right here, where it says mean, and standard

deviation, those you know, I teach and I think I understand that pretty well, but I wish I knew more behind all of it, then I would feel more comfortable saying that, you know, I understand it. (PreI, June)

When comparing distributions represented as histograms, June looked for “where most of the data is located” (PreI, June). She considered the shape of the distribution and the range. She tried to estimate the mean by mentally rearranging the data: “You don't know exactly what the numbers stand for, but in my mind, I try to kind of move little pieces here and there to figure out where the average would be” (PreI, June). She recognized that distributions can have the same center but different spreads. She mentioned that she was not clear on why sometimes the median is a better measure of center than the mean. She also mentioned looking for outliers as important. When comparing distributions with boxplots, she identified the median, the range, and compares corresponding quartiles. She seemed to understand that the density of a quartile varies inversely with its length. When comparing distributions as dotplots, June’s process was the same as it was for histograms.

When reasoning about variability, June was concerned with symmetry and normality as both were indicators of less variability to her. She also considered the range, with larger range indicating greater variability. There was no evidence that she viewed variability from the perspective of average distance from a central anchor point, but rather viewed symmetry with center and opposing data values “canceling each other out” (PreI, June) as with the mean. She appeared to understand significant difference as a big difference or one that depended on context. She did mention that the magnitude of the difference was relative to the range of the data, so that small differences could still be

statistically significant. She knew something about the relationship between significance and error and something about confidence, but she could not articulate her thoughts fully.

When asked about the experimental design question, June mentioned that she thought about this question after the pre-assessment. On pre-assessment Task 13 (Item 9), June wrote, “I really do not know” (Pre, June). During the interview, she mentioned:

You know, well how they set it up, I think that they should have two groups comparing people who live on a farm, and you know and how many allergies are present there. And then people who do not live on farms. And, I would think that you would have to look at like a longitudinal thing, because you've lived on a farm for like three or four days. You may not have had a chance to develop those allergies. I would think that you should maybe do this study like over time. And then they would have to have a way to measure the allergies and whether or not those allergies have anything to do with the environment of the farm. Because maybe they're allergic to like milk or eggs, or something you know what I mean, that's an allergy. But is that the type of allergy that is being discussed here? (PreI, June)

Clearly June had given some thought to the assessment item since the pre-assessment, but still, like others in this study, she did not attend to the formulation of a researchable question. When responding to the performance task prompt, June readily interpreted the data from the simulation and connected it to the problem. She was confident in her conclusion and this was especially surprising given that her confidence on the other interview tasks appeared not nearly so high.

June's general lack of confidence in her statistical knowledge was evident throughout the interview and substantiated by her self-assessed comfort-level with statistics of 1.57 out of 5 (see Table 9 in Chapter III). Though she spoke of shape, center, and spread, to some extent, her understanding did not appear to go beyond basic procedural and computational fluency. Her overall pre-assessment score of 1.40, with

demonstrated level of understanding of 2.25 for distribution, 1.00 for variability, and 0.67 for sampling distributions was consistent with her interview responses.

Callie. The pre-interview with Callie lasted approximately 41 minutes. Callie had been teaching high school mathematics for 18 years in an alternative school that was part of a relatively large rural school. She had taken one statistics course 20 years prior and had never taught a statistics course. She was unable to recall a single thing about her statistics course experience. She mentioned that statistics is not a strong thread in her high school and that she just teaches the basics of mean, median, and mode. Callie was challenging to engage with during the interview as she tended to answer with very short, often cryptic, responses.

When comparing distributions represented as histograms, Callie wanted to know the context, scale, counts, and range. She looked for similarities and differences. When comparing distributions as boxplots, she looked at the median, the range, and where the majority of the data lie. Her dotplot reasoning was nearly identical to that of histogram, although she did mention counting data values to make comparisons. When talking about variability, she seemed to use a personal definition referring to the closeness of data values to the center. Although her conception was not articulated completely, she explained that she consistently applied the same strategy when comparing pairs of distributions with respect to their variability. Her written responses for pre-assessment Tasks 10-12 (Item 9), suggested otherwise. In particular for Task 10, her response was “Group 1 because a larger percent of the group falls in the outliers;” for Task 11, she wrote, “Course 2 because the frequency does not gradually increase or decrease. It goes way up and then way down;” and for Task 12, “School B because very few students are

the mean height” (Pre, Callie). The written responses were interpreted as inconsistent; in particular, Task 10 and Task 12 were similar, but Task 11 potentially represented variability as “bumpiness.”

Callie’s view of significant difference was dominated by the idea that a large difference, presumably between means, implied significant difference. She seemed to want to have large amounts of data from which to make judgments and appeared to have a sense that the magnitude of the difference would depend on the range of the data. She did not remember any statistical way of determining significance. When responding to the prompt about experimental design (Item 9) she did not mention the importance of formulating a researchable question. Her written response to pre-assessment Item 9 was generic: “data collection method reasonable, contains only one variable, contains a control group, unbiased, accuracy of statistics, appropriateness of method used to determine results, conclusion & explanation” (Pre, Callie), but suggested that she had some understanding of experimental design. On the performance task, Callie was unable to associate the simulated data with the context of the question and respond to the question without a great deal of assistance. She acknowledged that she used very little technology in her school and did no work with simulations in mathematics classes.

It appeared that Callie was not particularly comfortable during the pre-interview. It may have been the case that her responses were constrained due to her lack of confidence or feeling of safety. Her overall pre-assessment score of 1.90, with demonstrated level of understanding of 1.75 for distribution, 1.67 for variability, and 2.33 for sampling distributions was generally consistent with her interview responses, with the

exception of the sampling distribution subset. Her sampling distribution score was surprisingly high given her interview responses.

David. The pre-interview with David lasted approximately 49 minutes. David had been teaching high school mathematics for eight years at a relatively large rural school, had taken one statistics course in college and another through the College Board for Advanced Placement Statistics teachers, and was the only self-proclaimed statistics teacher in the pool of interviewees. His pre-assessment content score was the second highest of all 56 teachers at 3.15 out of 4; his self-assessed comfort-level with statistics was 4.57 out of 5. He described his college course as having a few activities and a couple of computer labs, but mostly theory. He liked the course and found it easy.

When comparing distributions represented as histograms, David said he would compare shape, center, and spread and was competent in his explanation. Interestingly, his confidence was not so high when asked whether two distributions would be significantly different. A conjecture was that perhaps his understanding of significant difference was procedural or formulaic, in which case, he may not have been accustomed to making predictions from graphs. When comparing boxplots, which were his favorite representation, David considered the median, the IQR, and the range. He described how he could estimate the mean in a boxplot based upon the symmetry or lack of symmetry represented. When reasoning from dotplots, which David least preferred, he almost immediately imagined a boxplot representation of the distribution for comparison purposes. He was consistently challenged to estimate the mean or the median of a distribution that was not symmetric. Again, this may suggest a procedural or computational disposition of statistics. During his comparison of the dotplots from pre-

assessment Task 9 (Item 7), David predicted the differences in the means of the two distributions would not be significantly different; he also was uncertain about Item 1 with the histograms. In both cases, the differences in the means were highly significantly different, suggesting that his graphical connection with significant difference may not have been highly developed.

David claimed his reasoning about variability related to standard deviation; however his responses to pre-assessment Tasks 10-12 indicated he used a variety of strategies for correctly determining which of two distributions had more variability. For example, his explanation for Task 10 was, “Group 1, smaller size and no duplicate data values.” For Task 11, he wrote, “Group 2. This group has more variation in the center of the distribution.” For Task 12, “Group A (close) spread is higher” (Pre, David). David was the only one of the interviewees that referred to significantly different in a notably statistical way:

If you're talking about statistical significance, umm you're sort of contemplating whether a change or a difference in data could happen by itself. Just by random variation that will occur, or is it due to some other change that has for some other reason has caused the change in the data.
(PreI, David)

When responding to the experimental design prompt, David's written response was a generic response any statistics teacher might construct but it too was absent attention to the formulation of a researchable question. When prompted, David suggested several ways in which to formulate questions and he mentioned that his students are required to conduct a similar type of study as the final exam for his course. David interpreted and responded to the performance task simulation with confidence and was able to reason from the simulated data effectively.

David's interview and pre-assessment were notably stronger than other teachers in this study, presumably due to his experience teaching statistics. He appeared to have a solid grasp on shape and measures of center and spread; however, there was some evidence that his understanding may have been procedurally-dominated, especially his notion of variability and graphic interpretation of significantly different. His overall pre-assessment score of 3.15, with demonstrated level of understanding of 3.63 for distribution, 2.67 for variability, and 3.00 for sampling distributions was consistent with his interview responses.

Summary of Pre-Interviews

Data gathered from the pre-interview served to inform the construction of the HLT for the study and to triangulate the findings from the written pre-assessment. From the interviews, it appeared that many teachers in this study may have developed a very narrow and procedural view of statistics. There was little evidence from the interviews to suggest that teachers' understanding of statistical ideas went beyond that of descriptive statistics, and even the descriptive statistical knowledge appeared to be quite procedural. Many teachers confessed to being able to compute the mean, median, mode, range, and IQR, but were not confident in their ability to interpret the meanings of these values to speak about data. Teachers appeared to value graphical reasoning and desired students to learn to reason appropriately from graphs, but their own graphical reasoning was generally limited to surface features of representations and their interpretations were often problematic.

Teachers did appear to reason about distributions as opposed to individual data points. Their reasoning about variability suggested some prior knowledge with standard deviation, although that knowledge was not strong and did not appear to go beyond general awareness or ability to calculate values using the built-in feature on a graphing calculator. Generally, variability for these teachers seemed to be associated with the range or IQR and a variety of other colloquial meanings. The meaning of significant difference seemed connected to issues of context or magnitude of difference in a measure of center. Teachers appeared familiar with the language of “significantly different,” but only the A.P. Statistics teacher could articulate an interpretation which approximated a statistical perspective.

None of the teachers in the interview suggested the need to formulate a researchable question for a given experimental design task and their suggestions for a high quality student response were quite generic. With respect to the simulation task, teachers were generally not able to connect the simulated data, generated under a given assumption, to the task at hand. Teachers discussed providing either very little or no experience for their students working with simulations, so it is plausible this type of task was not familiar. Because statistics appeared to occupy a very small part of the mathematics curriculum for these teachers, perhaps reasoning from simulations was particularly unfamiliar.

Interview teachers’ average initial level of understanding of comparing distributions, as measured by the content pre-assessment (1.89), was consistent with the interview data. The translation from the theoretical framework presented in Chapter III of teachers’ understanding of comparing distributions is considered emerging distributional

(first holistic view of the data; informal qualitative descriptors of the data, along with basic summary statistics, are used to describe two datasets. Teachers begin to understand the difficulty in creating measurable conjectures, but are unable to successfully resolve the conflict and show frustration in attempting to write an appropriate conjecture. Variability, while acknowledged, is not understood beyond a descriptive level), with respect to variability, they demonstrated partial recognition of variation (put ideas in context, tendency to focus on single aspects and neglect others), and for sampling distributions, they are at the verbal reasoning stage (have a verbal understanding of sampling distributions and implications of the Central Limit Theorem, but cannot apply this to the actual behavior of sample means in repeated samples). The pre-interviews strongly corroborated the performance on the written pre-assessment and informed the development of the HLT described in Chapter IV. Pre-assessment scores for the set of teachers interviewed, presented in Table 47, confirm that the group of teachers interviewed was remarkably similar at pre-assessment to the entire group. The mean score for the nine interview teachers on the pre-assessment was identical to the 1.89 for the entire sample of 56 teachers.

Post-Interviews: Summary and Emerging Themes

Post-interviews with the same set of teachers who participated in pre-interviews, were conducted eight days following the end of the professional development program. The purpose of the post-interviews was to further inform the study regarding teachers' understanding of comparing distributions as well as to hear from them about the ways in which the professional development program, especially the use of resampling techniques and dynamic statistical tools helped to shape their understanding.

Table 47

Interview Teachers' Pre-Assessment Scores by Item

Name	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Average
Cameron	2	1	4	2	1.5	1	2	2	2	1	1.85
Jessy	2	3	4	1	0.5	1	2	2	2	1	1.85
Lorraine	2	1	4	1	2	2	3	2	2	1	2.00
Jaylee	2	0	1	2	3	0	2	0	0	0	1.00
Sasha	2	1	4	2	1	1	2	2	1	1	1.70
Jordan	3	1	4	3	4	1	1	2	2	1	2.20
June	3	1	1	1	3	2	2	1	0	0	1.40
Callie	2	2	1	1	2	3	2	2	2	2	1.90
David	4	2	4	4	3.5	3	3	2	2	4	3.15
<i>M</i>	2.44	1.33	3.00	1.89	2.28	1.56	2.11	1.67	1.44	1.22	1.89
<i>Mdn</i>	2	1	4	2	2	1	1	2	2	1	1.85
<i>SD</i>	0.73	0.87	1.50	1.05	1.18	1.01	0.60	0.71	0.88	1.20	0.59

Table 48 contains the post-assessment scores for the interview subset of nine teachers. As can be seen by comparing the results from Table 48 with those from Table 47, a positive shift in scores can be seen for individuals by item and overall from pre- to post-assessment. Table 49 compares the pre-, post-, and gain-scores for interviewed teachers, non-interviewed teachers, and all 56 teachers in the study.

Table 48

Interview Teachers' Post-Assessment Scores by Item

Name	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Average
Cameron	3	4	4	4	3	3	2	3	1	4	3.10
Jessy	3	2	4	4	3.5	4	4	3	2	4	3.35
Lorraine	3	2	4	3	2.5	3	3	2	4	2	2.85
Jaylee	2	4	4	3	3.5	3	3	3	2	4	3.15
Sasha	2	4	4	3	3	3	2	2	2	2	2.70
Jordan	2	4	4	3	3.5	4	3	3	2	4	3.25
June	4	3	4	3	3.5	4	3	4	2	4	3.45
Callie	3	3	1	4	3.5	4	3	3	4	4	3.25
David	4	4	4	4	3	4	3	4	3	3	3.60
<i>M</i>	2.89	3.33	3.67	3.44	3.22	3.56	2.89	3.00	2.44	3.44	3.19
<i>Mdn</i>	3	4	4	3	3.5	4	3	3	2	4	3.25
<i>SD</i>	0.78	0.87	1.00	0.52	0.36	0.53	0.60	0.71	1.01	0.88	0.28

Table 49

Interviewed Teachers, Non-Interviewed Teachers, and All 56 Teachers: Comparison from Pre- to Post- and Gain-Scores

	Pre-assessment			Post-assessment			Gain-scores		
	Interview Subset	Non-Interview	All	Interview Subset	Non-Interview	All	Interview Subset	Non-Interview	All
Item 1	2.44	2.17	2.21	2.89	2.66	2.70	0.45	0.49	0.48
Item 2	1.33	1.91	1.82	3.33	2.53	2.66	2.00	0.62	0.84
Item 3	3.00	2.21	2.34	3.67	3.23	3.30	0.67	1.02	0.96
Item 4	1.89	1.40	1.48	3.44	3.13	3.18	1.55	1.73	1.70
Item 5	2.28	2.51	2.47	3.22	2.99	3.03	0.94	0.48	0.55
Item 6	1.56	1.83	1.79	3.56	3.06	3.14	2.00	1.24	1.36
Item 7	2.11	2.03	2.04	2.89	2.59	2.64	0.78	0.57	0.60
Item 8	1.67	1.90	1.86	3.00	2.46	2.55	1.33	0.58	0.70
Item 9	1.44	1.70	1.66	2.44	2.30	2.32	1.00	0.59	0.66
Item 10	1.22	1.28	1.27	3.44	2.07	2.29	1.22	0.98	1.02
Average	1.89	1.89	1.89	3.19	2.70	2.78	1.29	0.81	0.89

Table 49 illustrates the relatively larger gains for the interview teachers from pre- to post-assessment in comparison to the non-interview teachers. The interview teachers outperformed the non-interview teachers on every item on the post-assessment. A possible explanation for the increased performance at post-assessment for the interview teachers may be that because they anticipated the post-interview, perhaps their “level of concern” was higher than teachers who were not participating in interviews. It is also possible that the pre-interview experience sensitized the teachers to issues seen as important by the researcher. A third plausible conjecture is that teachers who were willing to be interviewed for a study like this may be teachers who are particularly motivated to learn. In any case, an interesting hypothesis coming from these findings is that the use of pre- and post-interviews as part of a professional development intervention may support improved learning on the part of teachers.

The post-interviews generally confirmed that teachers’ understanding of distribution, variability, and sampling distributions had improved, leading to a general

improvement in their understanding of *comparing distributions*. Aside from the improvement in statistical content knowledge as measured by the written assessment and confirmed during the post-interview, analysis of the post-interviews revealed a number of additional teacher changes: (1) teachers' willingness and ability to speak with statistical language improved markedly; (2) teachers' improved understanding of variability tended to include the use of standard deviation or an average distance from a central anchor point perspective and take into account shapes of distributions and sample size; (3) teachers' understanding of statistically significant differences improved beyond simply describing the magnitude of a difference or a difference based on a particular context, to include sensitivity to sample size, variation, use of the randomization test to determine p -values, differences in locations of clusters of data values, and coordination of histogram or dotplot representations with boxplot representations; (4) teachers' facility with dynamic statistical software increased from none to that of quite proficient; (5) teachers' understanding of sampling distributions appeared to be stronger following the professional development. Sample illustrations from the post-interviews are presented next. The samples were selected for either their representative quality or to serve as contrasts from which to compare. In all, more than 400 pages of post-interview transcripts were analyzed and the pieces presented are those most essential for helping to answer the research questions for this study.

Teachers' Improved Use of Statistical Language in the Context of "Significant Difference"

To ascertain how language changed from pre- to post-interview, it is instructional to look at responses to questions asked in both interviews. "What does significantly

different mean?” is an example of an interview prompt common to both interviews.

David is the only example of a teacher for whom the definition of statistical significance appeared well-developed at the pre-interview. When asked about it, he responded,

If you're talking about statistical significance, umm you're sort of contemplating whether a change or a difference in data could happen by itself. Just by random variation that will occur, or is it due to some other change that has for some other reason has caused the change in the data . . . you would probably have to look at the rest of the distribution and see how spread out that is. I mean there are different cases where point one and whatever units you happen to be measuring in, is significant and where it isn't. So it depends on how it's relative to the rest of the distribution. (PreI, David)

He went on to discuss boxplot representations and to suggest that if the IQRs overlapped, then the mean or medians of two distributions were not likely significantly different. At the post-interview, David's conception remained intact, but offered additional insight

(PostI, David):

David: *The probability of finding that difference by the natural variation that will occur anyway, is . . . are getting a . . . a difference that large or larger strictly by that random variation or that natural variation is low . . . below 5%.*

SM: *How do you feel about how the randomization test helps flesh that out?*

David: *I think it . . . it hits the fact very clearly that you're making an assumption when you pull the data together that these things come from the same pool. So there is no difference. That's . . . that's the null hypothesis right there. There is no difference, so let's operate on that assumption and then, if that's the condition that we're under, let's see what happens. We know what the mean from whatever sample is. Let's say if . . . if they're together, how often is that mean going to occur? And that's our sampling distribution, so you can see how far away it is from that natural bell curve, or how close it is and how, yeah, that happens a lot or yeah, that doesn't happen very often. So it's a pretty good way to do it.*

David went on to explain that he uses t -tests and z -tests in his classes and that his students often struggle with the logic of hypothesis testing and calculating p -values from charts.

He had not experienced the use of the randomization test prior to the professional development program and thought it might be a good vehicle to assist students' learning.

Cameron's post-interview response provides a contrast because his pre-interview response suggested reasoning about significant difference based upon the magnitude of the difference:

*Well, the . . . the group . . . the group's scores are much . . . as a group, are much greater . . . much better than another group scores.
[So what makes it significant? How do you know?]
Because that's just being different . . . Large enough to make a . . . a di—
make a substantial difference. (PreI, Cameron)*

At the post-interview, his response suggested a change in thinking about statistical significance to include a relation to p -value:

Well, for me, what I brought out of the . . . out of the . . . one of the main ideas that I took out of the sessions was statistical . . . significance. Significantly sta—statistically significant was its relation to the p -value. And . . . and I . . . as we . . . as we just did [referring to a performance task part of the interview], I went right to the p -value, which maybe is not the . . . sometimes it's not the best thing to do, but it's an indicator that the statistics are telling us something that is very important and it's just not something that's random or happening out of, by chance, but the statistics are actually communicating to us something . . . a . . . a legitimate conclusion. Now, legitimate conclusion meaning something that's highly likely, not that it's 100% sure, but it . . . we have strong indicators to suggest that that is true, that our hypothesis is true. So when we say statistically significant, I'm thinking right away, p -value, although that's not the only indicator. There are other indicators, but for me, because I laid . . . I laid a hold of the p -value in the sessions, it relates directly to . . . to that. (PostI, Cameron)

He went on to describe the physical simulation of the randomization test and to adequately describe how the process is used to come to a conclusion. The experience with the randomization test appeared to have impacted Cameron's idea of statistical significance as seen through his language of random, by chance, and p -value.

According to Jessy, statistical significance meant

First of all, I think it depends on what level . . . what p-value you're willing to . . . if you're doing drug trials, then you would want . . . you would want it to be so rare that it happened by random in order for you to conclude, especially when lives are on the line. But if it's just comparing something where you might accept a p-value like 0.10, if we, you know, you had . . . like you . . . the example we had, where you had to choose the . . . this paper over this paper for your . . . your drop, you don't really have to be, I think, as sure of yourself, because you have to choose one regardless. And so I think the p-value comes into play when you are deciding if it's statistically significant. So I would say, and the event is statistically significant if the difference in the mean occurs so rarely by . . . by chance, that you can attribute your treatment causing the event. For an example—like for an example, if . . . if by randomization table, it shows that this difference occurs one time out of a thousand, and your treatment also caused that, then I would be pretty confident that it was my treatment that caused it because it's so rare. So I would say that my data was statistically significant. If the randomization table showed that 40% of the data was to the right of that red line, which shows 40% of the time it happens by chance, then I would not think that my result was that statistically significant. (PostI, Jessy)

Though language use varied from teacher to teacher, all of the teachers communicated an improved understanding of significant difference at the time of the interview. For eight of the teachers, that change in understanding appeared to be heavily influenced by teachers' work with the randomization test. One teacher, June, was confused during the interview about the randomization test but she could reason confidently comparing two distributions based upon proportional clusters of data values related to differences in location. Pieces of June's interview are presented in considerable detail in the following section.

Randomization Testing and the Use of Fathom2 and CPMP-Tools

When working on the performance task part of the interview, June compared the two distributions in the task by looking at the proportion of the data that clustered above and below a certain point. She was confident in her comparisons and conjectured that the

difference in means between the two groups would be significant. She asked if she should conduct the randomization test like we did in class. The following transcript conveys June's confusion and the way in which the interviewer was trying to help her reactivate her experiences from the professional development (PostI, June):

- SM: *Would you be comfortable doing that?*
June: *I think so. I have to tell you, I'm kind of forgetting everything we did in the class. There's so much other things happening in my life . . .*
SM: *Sure.*
June: *. . . but I think I could probably get through it.*
SM: *Okay.*
June: *So what I'm thinking is . . . I'm a little confused about that randomization test because I don't remember having only two different groups when we did that. I remember having like one set of data. And we, you know, scrambled it.*
SM: *Yeah.*
June: *And we want to see if it was likely that that group of numbers would show up, or how likely it was that that group of numbers would show up. And then, you determined whether or not something was happening based on the probability that that would have happened, otherwise. So with this would I take just one of the groups and do that?*
SM: *Oh, okay, so . . .*
June: *Or am I being totally on the wrong track here?*
SM: *Well, no, you're . . . you've got actually, you're on the . . . you're on a really good track, but there seems like there's one missing link. Remember when we first started out, the very first day, when we did the . . . the Orbital Express . . .*
June: *Um-hum, that's what I'm thinking of, yeah.*
SM: *. . . and we had . . . we had those two kinds of papers? And we dropped . . . we did seven drops of one, and we did seven drops with another. So we had those measurements and we wrote them on two different colored cards, which kind of would be like this situation right now. And then, what did we do with those cards?*
June: *Oh, mixed them together. Okay.*
SM: *Yeah, we mixed them together.*
June: *Okay.*
SM: *So now, what is that . . . why did we mix them together?*
June: *Well, you're assuming that there's no difference.*
SM: *Okay.*
June: *Okay, so that's why you're allowed to mix them together because supposedly they're all the same [from the same population].*
SM: *So we mixed them together and then what'd we do?*
June: *We shuffled them up.*

SM: *Shuffled them up.*
 June: *And then, I'll redistribute them in two groups.*
 SM: *Into two groups, just like they were originally.*
 June: *Okay, okay.*
 SM: *And then, what did we do?*
 June: *Um . . . [long pause]*
 SM: *What were we trying to measure?*
 June: *The probability that each of those groups would have occurred naturally?*
 SM: *What about each of those groups? We did something with this group.*
 June: *Oh, the difference in the means?*
 SM: *Yeah, we found the mean [group 1].*
 June: *Okay.*
 SM: *We found the mean [group 2]. We found the difference.*
 June: *Okay.*
 SM: *And that was a number. And it usually didn't match what our original difference was.*
 June: *Okay.*
 SM: *And so we . . . that's what we accumulated. We did that process a lot of times.*
 June: *Okay.*
 SM: *And then we . . . so we got a whole bunch of those.*
 June: *So we had a . . . a . . . as many as we wanted.*
 SM: *Exactly.*
 June: *Like 500.*
 SM: *Yeah.*
 June: *The difference of the means.*
 SM: *Exactly.*
 June: *Okay.*
 SM: *And then, we compared our true difference . . .*
 June: *Okay.*
 SM: *. . . to this distribution of differences.*
 June: *Okay.*
 SM: *And I think that's what you were talking about.*
 June: *And . . . and that . . . that graph that we created after that, then is that what we looked at, like the . . . the probability and if there was only a tiny little bit out here . . .*
 SM: *Uh-huh.*
 June: *. . . then that meant that there was a small . . . let me think.*
 SM: *Yes.*
 June: *A small chance that this would have occurred? Yeah.*

The transcript illustrates that for this teacher, understanding of the randomization test procedure for comparing means of two distributions was clearly still developing. With

prompting and reference to a specific physical simulation activity from the professional development program, she was able to reconstruct most of the process. She appeared to get hung up on the creation of a measure to compute the difference in means of the two distributions and perhaps in thinking about what the randomization distribution actually represented.

What is particularly peculiar about this case is when June moved to the *Fathom2* environment, she demonstrated that she could build the mechanism to conduct the randomization test. She stumbled a little on the creation of the formula for the differences in the means, almost the identical place she stumbled in the example previously mentioned, but she generated the stacked collection, the scrambled collection, and the collection of measures representing the randomization distribution. She spoke through the entire process and indicated amazingly strong facility with the software and the randomization test process. She moved easily between the hierarchy of collections and a variety of graphical and tabular representations. She navigated well with the inspector, a task which is often challenging for beginning *Fathom2* users. During the interview, she mentioned that after doing the initial tours for the software, she did not actually drive the computer during the professional development program (PostI, June).

- June: *I didn't actually ever have a chance to really play with that, um, in the class. The first day of that tutorial thing [Fathom2 tours] . . .*
- SM: *Yeah.*
- June: *. . . I went in the computer lab and worked through . . . I think we had to do like one, two, and six, and I got through two. And I really followed it very well when I was working in a group, but I didn't actually have a chance to do it.*
- SM: *You weren't driving the computer?*
- June: *No.*
- SM: *What? You are doing remarkably well.*
- June: *I'm really good with this stuff.*
- SM: *Okay.*

June: *You know, it's just I don't remember things is the problem right now, and I'm not used to this [the laptop computer].*

As the interview continued, June eagerly demonstrated more capabilities of the software as she continued to explore multiple graphical representations and plotted values on graphs and made various comparisons. Given that she did not navigate in *Fathom2* beyond the tours during the professional development program, June is a prime example of a teacher able to utilize dynamic statistical software to augment and support her statistical understanding. As the interview went on, it became more and more apparent that June was able to use the software to explore myriad relationships with data and her communication of ideas continued to improve. It was as if she was remembering things by virtue of navigating in *Fathom2* and becoming increasingly confident in her reasoning as the interview progressed. During the interview, June eventually convincingly demonstrated a powerful understanding of the randomization test process for comparing means or medians of two distributions along with many connections to normal distributions, standard deviations, shapes of distributions, measures of centers, sampling distributions, and context.

Three other teachers, Callie, Jaylee, and Lorraine, demonstrated a similar particular affinity for and prowess when working with *Fathom2*. Lorraine, like June, apparently did not have the opportunity to navigate with the computer during the session (due to her partner), but was able to do so during the interview. To the surprise and delight of the researcher, Jaylee, with the lowest pre-assessment content score, proficiently used *Fathom2* to conduct a randomization test. Her informal comparisons of two distributions prior to using the technology indicated a sensitivity to sample size, shape, center, and spread. Her use of *Fathom2* was nothing short of exceptional. Her

post-assessment score showed the greatest gains of all 56 teachers. Callie was meek and hesitant at the time of pre-interview; at the post-interview, she was in command of what she was doing and saying and demonstrated an uncanny level of confidence with and without the software. The use of the *Fathom2* software and the computer is likely an important factor in the gains in statistical understanding made by her and for the teachers as a whole.

Of the nine teachers, when given the opportunity to use a dynamic statistical tool for a statistical investigation, six utilized *Fathom2*, two utilized *CPMP-Tools*, and one declined to use the software but confidently conducted the analysis. A likely explanation for the predominant use of *Fathom2* was the relatively stronger familiarity with the software because of the time of its use during the professional development program than with the briefly used *CPMP-Tools*. Still, given the complexity of the construction of the randomization test mechanism in *Fathom2*, it was surprising that two-thirds of the teachers would elect to build the mechanism from scratch to conduct an investigation. Ironically, the teacher with the strongest initial content knowledge as measured by the pre-assessment was the most reluctant to use any software during the interview. The use of *CPMP-Tools* and *Fathom2* software clearly played a role in teachers' understanding of comparing distributions, whether it was directly used or even just observed as part of a group. Because of the cases where teachers did not actually use the software during professional development, but were able to effectively use it during the interview, it is likely that the design of the software is intuitive for the novice learner.

Makar (2004) studied teachers' behaviors while learning to use technology (more specifically, *Fathom*) to conduct investigations. She categorized teachers as Wanderers,

Wonderers, or Answerers as described in Table 50. During the post-interview in the present study, teachers were asked to select the category they felt best represented how they use technology (*Fathom2*) to conduct investigations. As Table 50 illustrates, only one of these teachers considered himself a Wanderer and only as a learner, not as a teacher. Six of nine teachers considered themselves Wonderers; two of nine considered themselves Answerers. Cameron categorized his use of technology as dependent upon his role. As a learner, he felt he was a 1/5 Wanderer, 3/5 Wonderer, and 1/5 Answerer, but as a teacher, he felt he was an Answerer. During the post-interview he said,

I . . . I just can't wander in class [laughter]. I . . . I've got to know the . . . if . . . if I'm teaching . . . and maybe this is . . . I mean, it is a fault of mine. I've realized that as a teacher, I have a hard time wandering in class. And I've worked on that. This . . . this in-service has helped me to think introspectively on that, and maybe I should do some more wandering, but because class time, there's so little of it . . . with the stupid things that are going on in our public schools, when I teach, I have to have the answers. And so I'm teaching as a control freak sometimes, leading, pushing in that direction, and so I don't have time to wander, which is . . . which is . . . I hate that. (PostI, Cameron)

Sasha said,

I would put myself with the Wonderers. Because I think although I want the answer, like the Answerers do, I want to know what's going to happen if I do other things. I don't want just the answer, I want more information about what it is that I'm doing and what it is that I'm looking at. (PostI, Sasha)

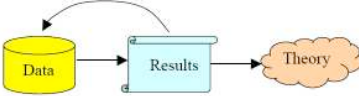
Jordan, a self-proclaimed Answerer, said,

Hm, I'm an Answerer. Find the answer and say, this is . . . Yeah. I want to find the answer and that . . . that's what I . . . Like I have a question and I'm searching for the answer. I want to get to the end. I don't . . . I don't wander at all. I'm not a Wanderer or a Wonderer. I want to know the answer. I'm definitely this one. (PostI, Jordan)

What feels striking about these teachers' self-characterizations is that really none of them considered themselves to be Wanderers. In Makar's (2004)

Table 50

Teachers' Self-Characterizations of Their Use of Technology

Teachers responding to, "How do you use technology (<i>Fathom2</i>) to conduct investigations?"	Description of Category (Makar, 2004)
Cameron (as learner)	<u>Wanderers</u>—Use data to look for a theory. Spend a good deal of time “wandering” through various analyses that are not necessarily directly connected to a conjecture, but hopeful that something will jump out that can be tied back to a conjecture. 
Callie David Jessy Lorraine Sasha June Cameron (as learner)	<u>Wonderers</u>—Lead by “I wonder” questions. Generally create a theory and then use the data to test their theory. 
Jordan Jaylee Cameron (as teacher)	<u>Answerers</u>—Go into an investigation with a theory like the Wonderers, but without generating “I wonder” questions during investigation. Search for a particular piece of evidence to support or refute their conjecture and then directly state their conclusion. 

study, 8 of 17 pre-service teachers were categorized as Wanderers (5-Wonderers, 4-Answerers). It is possible that teachers' views of their use of technology for conducting investigations, as in the present study, may not match a researcher's view, as in Makar's study, or maybe pre-service teachers are qualitatively different in this respect than experienced teachers. It is not known whether the characterization selected by teachers was attributable to their work during the professional development or a more general phenomenon. Regardless of the teachers' categorizations, a strong affinity to finding answers permeated the teacher interviews. On the other hand, given that seven of nine teachers in the present study characterized themselves as Wonderers, an hypothesis might

be that high school mathematics teachers tend to be quite goal oriented, but perhaps curious as well.

Possible Misconceptions or Language Issues

One of the things that surfaced through the post-interviews that remained camouflaged throughout the other analyses for the study was teachers' struggle with language potentially representing misconceptions. During the interviews, there were three distinct uses of language that signaled possible difficulties. The first example refers to a possible confusion between samples and populations. It also contains a mention of the probabilistic idea of sample space. It is plausible that the teacher was referring to sample size.

Without technology, I...I would be concerned with the number of the se—if I'm going to compare them, the sample populations are so dissimilar in size, 47 as compared to 187, I...I'm wondering if that maybe is sufficient...a sufficient sample space to...to be able to compare fully.
(PostI, Cameron)

A second example illustrates possible confusion between samples size and number of trials. At the post-interview when talking about Item 10 on the post-assessment (related to the interpretation of sampling distributions and the Central Limit Theorem), Lorraine verbally responded in the following way:

I didn't know if 500 samples of 16 would be big enough to make it because that was kind of like . . . well, we were at, I guess, about 500 samples when it got tall and skinny. And this one's still at 500 samples, but they were only samples of four, so I figured it would still start to be getting normal, but still be kind of spread out. (PostI, Lorraine)

Lorraine's written answers to post-assessment Item 10 suggested she could select the appropriate graphical representations and she answered 4 of 7 parts of the item correctly, but her incorrect responses suggested confusion about relative variability.

Referring to the same post-assessment item, Jessy's response included the following language that suggested she was possibly confusing sample size and number of trials as well. She appeared to understand that the variability of the sampling distribution of the mean will become approximately normal with less variability as sample size increases, but her language, especially in the last sentence is still problematic.

And I remember the time we sat on our computer and everybody had a different population, and we kept doing them, and they kept getting narrower and taller. So I was thinking that A went with this one because the sample size was larger, so I thought it would be more accurate to getting more . . . smaller or narrower. And the more you do, the more, closer it should resemble that tall, skinny, normal curve. And I knew this one should . . . sort of make a normal curve, but since it's only a sample size of four, I thought that maybe it would have more variability because you didn't take as large of a sample . . . We all had normal curves, but the more . . . the more times we took samples, the narrower and wi—and taller it became. (PostI, Jessy)

Jessy's written answers on post-assessment Item 10 were all correct but her use language was problematic and suggested she may have been still wrestling with the difference between samples and sampling distributions. It is also possible that the confusion may have just been a semantic issue.

Finally, another response heard in the post-interviews and seen on some of the written assessments was related to associating normal-looking distributions with relatively less variability as compared to non-normal-looking distributions with relatively greater variability, regardless of scale. For example when discussing her idea on variability, though she referred to standard deviation in her explanation, Lorraine also said,

. . . and I saw it somewhere in my head, it said that, the more normal the distribution, the less the variability. (PostI, Lorraine)

Though the proportions of responses like those above were low, they certainly have implications for future iterations of design work in this area.

Teachers' Overall Change in Understanding of Comparing Distributions

Teachers were asked in what ways their understanding of comparing distributions had changed during the sessions. In response, the things teachers mentioned or demonstrated included connections between graphical representations and summary statistics; a much better understanding of multiple measures of variability, but especially standard deviation; a completely new understanding of the randomization test to compare distributions; the relationship between sample size and standard error; aspects of experimental design; and how inadequate their previous ideas of comparing distributions may have been. Table 51 illustrates numerical post- and gain-scores from the written assessments for sub-ideas and overall as well as the length of time of the pre- and post-interviews for comparison.

Table 51

Summary of Teachers' Interview Times and Change in Understanding on Post-Assessment by Sub-Idea and Overall

Teacher	Site	Interview Duration (min.) Pre→Post	Distribution (gain)	Variability (gain)	Sampling Distribution (gain)	Overall Post- Assessment (gain)
Cameron	1	53→56	3.00 (1.13)	3.67 (1.33)	2.67 (1.33)	3.10 (1.25)
Jessy	1	53→49	3.63 (2.25)	3.00 (0.00)	3.33 (2.00)	3.35 (1.50)
Lorraine	1	35→48	2.88 (0.88)	2.67 (0.33)	1.67 (1.33)	2.85 (0.85)
Jaylee	2	36→44	2.88 (0.63)	3.67 (3.33)	3.00 (3.00)	3.15 (2.15)
Sasha	2	47→57	2.50 (0.75)	3.33 (1.00)	2.33 (1.33)	2.70 (1.00)
Jordan	2	39→47	2.88 (0.13)	3.67 (1.33)	3.33 (2.00)	3.25 (1.05)
June	3	39→84	3.38 (1.13)	3.67 (2.67)	3.33 (2.67)	3.45 (2.05)
Callie	3	41→53	3.38 (1.63)	2.33 (0.67)	4.00 (1.67)	3.25 (1.35)
David	3	49→41	3.50 (-0.13)	4.00 (1.33)	3.33 (0.33)	3.60 (0.45)

For completeness, Table 52 presents the change in teachers' self-reported comfort-level with statistics and their pre-, post-, and gain-scores for the written assessments including all ten items and then with the subset of eight parallel items.

Table 52

Summary of Interview Teachers' Content Knowledge and Self-Reported Comfort Level with Statistics Pre- and Post-Intervention

Name	Site	# Stats Courses Taken	Pretest (10 Items)	Posttest (10 Items)	Gain (10 Items)
		<i>Comfort-level— Pre → Post</i>	<i>Pretest (8 Items)</i>	<i>Posttest (8 Items)</i>	<i>Gain (8 Items)</i>
Cameron	1	1 <i>1.71 → 1.93</i>	1.85 <i>1.94</i>	3.10 <i>3.38</i>	1.25 <i>1.44</i>
Jessy	1	1 <i>2.29 → 2.57</i>	1.85 <i>1.94</i>	3.35 <i>3.44</i>	1.50 <i>1.50</i>
Lorraine	1	1 <i>1.57 → 1.86</i>	2.00 <i>2.00</i>	2.85 <i>2.69</i>	0.85 <i>0.69</i>
Jaylee	2	2 <i>1.42 → 3.43</i>	1.00 <i>1.25</i>	3.15 <i>3.31</i>	2.15 <i>2.06</i>
Sasha	2	1 <i>2.14 → 3.43</i>	1.70 <i>1.88</i>	2.70 <i>2.75</i>	1.00 <i>0.88</i>
Jordan	2	1 <i>2.83 → 4.14</i>	2.20 <i>2.38</i>	3.25 <i>3.31</i>	1.05 <i>0.94</i>
June	3	1 <i>1.57 → 3.57</i>	1.40 <i>1.50</i>	3.45 <i>3.56</i>	2.05 <i>2.06</i>
Callie	3	1 <i>1.57 → 3.29</i>	1.90 <i>1.75</i>	3.25 <i>3.06</i>	1.35 <i>1.31</i>
David	3	2 <i>4.57 → 4.71</i>	3.15 <i>3.31</i>	3.60 <i>3.63</i>	0.45 <i>0.31</i>
AVERAGE <i>n_{interview} = 9</i>		1.22 <i>2.19 → 3.21</i>	1.89 <i>1.99</i>	3.19 <i>3.24</i>	1.29 <i>1.24</i>
AVERAGE <i>n_{all} = 56</i>		1.18 <i>2.22 → 3.17</i>	1.89 <i>1.94</i>	2.78 <i>2.79</i>	0.89 <i>0.86</i>

The teachers interviewed for this study made gains in content knowledge as measured by the written assessments and communicated during the interviews. These teachers also made impressive gains in their developing facility with dynamic statistical technology as demonstrated during the interviews. Some issues of problematic language usage and potential misconceptions surfaced during the interviews that were not seen readily from other data sources. The data from the interviews reinforced and augmented

what emerged from the content assessments, and in the following section, will be shown to be consistent with analyses of both the written reflections and video from the professional development.

Written Reflection Results

At five distinct times during the four-day professional development program, teachers were asked to respond in writing to a set of reflection prompts (Reflection 1, Reflection 2, Reflection 3, Reflection 4, and Reflection 5). The prompts were designed as formative assessments to provide ongoing feedback to guide instructional decisions as well as to serve as artifacts for use during the retrospective analysis. The five sets of written reflection questions included 24 separate prompts and can be found in Appendix E. All of the reflections reported are based upon $n = 56$ teachers.

The organization of this section is generally chronological with Reflection 1 responses followed by Reflection 2 responses and so on. Multiple prompts provided in a reflection setting are coded as R2-A, R2-B, R2-C to represent Reflection 2, items A, B, and C. Generally, Reflection 1 responses are reported together, then Reflection 2 responses, etc. Some prompts were repeated on multiple reflections in order to track change on some ideas over time. The repeat reflection responses will generally be presented collectively with the reflection results from the final reporting opportunity for a given idea. There are exceptions to this rule; for example, when several prompts on one reflection refer to a similar idea and a follow-up prompt on a later reflection, all of these responses are reported in the same section.

Reflection 1

Reflection 1 occurred following the completion of Activity 1.4 and contained the following writing prompt:

Based on today's experiences, please describe ways in which your understanding of some statistical ideas has changed (what do you understand better now than you did this morning--please explain?)

Qualitative responses were first coded with $-$, 0 or $+$ to reflect teachers' reflections as negative, neutral, or positive (Table 53). Subsequently, each response was coded according to the topic(s) addressed. Additional codes were constructed for unusual or particularly insightful responses. Overwhelmingly, the responses were positive with 53 of 56 (95%) noting specific statistical ideas for which they had a better understanding after the first day's activities. By a large majority, the most cited statistical idea with improved understanding for teachers was that of standard deviation. Many teachers acknowledged that they had limited or no prior understanding of the reason for dividing by $n-1$ for sample variance versus by n for population variance. Many teachers admitted to lacking knowledge of the difference between a statistic and a parameter. The use of the symbols μ , $\bar{x}$, σ , s_x became clear for many during the first day. As Table 54 illustrates, the majority of responses to this reflection prompt suggest an improved understanding of issues related to variability and its interpretation. A second major clustering of big ideas with self-reported better understanding is related to improved ability to match distributions to context or summary statistics and to use representations of histograms and boxplots simultaneously to compare distributions. The frequency of the coded responses was used to sort the list in descending order. A number of the categories may have been able to be collapsed; however, the choice was made to maintain the longer list as

illustrative of the variety of responses seen on this prompt. One thing suggested from these reflections is the verification of the landscape of big ideas that were investigated in roughly five hours of professional development. Another observation is that the same or similar professional development experiences may impact individuals' understanding in different ways. This notion will continue to surface as additional reflection prompts are analyzed.

Table 53

Reflection 1: Nature of Teachers' Responses by Site

	Site 1	Site 2	Site 3	Total
–	0	0	0	0
0	2	1	0	3
+	29	8	16	53
Total	31	9	16	56

Table 54

Reflection 1: Frequency of Big Ideas for which Teachers Reported Improved Understanding

Improved understanding of statistical idea	Frequency	Improved understanding of statistical idea	Frequency
Standard deviation	39	Context	3
$n-1$	23	Groupwork (not really statistical)	2
Variance	14	MAD (mean absolute deviation from the median)	2
Population vs. sample	13	Outlier definition	2
Mean vs. median	12	Scales	2
σ vs. s_x	10	Stat-math relationship	2
Matching	9	Summary statistics	2
Skewed	8	Bar graphs vs. histograms	1
Different estimators	5	Bias	1
Fathom	5	Central tendency	1
Representations	5	Definitions	1
Statistic vs. parameter	5	Experimental design	1
Calculator	4	Interpret data	1
Normal distribution percentages	4	Mean as balance point	1
Shape	4	Mean related to variance	1
μ vs. $\bar{x}$	3	Normal distribution shape	1
Activities	3	Probabilistic answers	1
Boxplots	3	Total	194

Reflection 2

Reflection 2 occurred following the completion of Activity 2.3 and was mainly concerned with teachers' understanding of the randomization test that had been introduced and begun to be developed. Six prompts on R2 included:

- R2-A In your own words, describe what you have learned about the randomization test.
- R2-B What is it useful for?
- R2-C What does it help you to figure out?
- R2-D What questions do you have about the randomization test?
- R2-E On a scale from 1 to 10, rate the ease of use of *CPMP-Tools* to conduct the randomization test. (1 is low; 10 is high)
- R2-F On a scale from 1 to 10, rate your current understanding of the randomization test.

Parts A – D are analyzed next and the discussion will include responses to R5-D from Day 4 because of its connection to this content.

- R5-D What questions do you still have about the randomization test?

The results of R2-E will conclude this section. Discussion of R2-F is included with the Reflection 5 discussion along with the other three prompts of this type.

R2-A through D plus R5-D

After reading through the responses to these prompts for each person, a code was established indicating whether their articulated understanding was relatively high (H), medium (M), or low (L). The medium category was further broken into M+, M, or M– to indicate relative understanding. Responses generally indicated at least a developing sense of the randomization test procedure. Sample responses for each category are provided in Table 55 along with the frequency of responses in each category. The majority of

teachers indicated a medium to high initial understanding at this point in the professional development program.

Table 55

Reflection 2: Categorization of Responses for Parts A, B, C, and D

Categorization Level		Representative Response at Level	Code #	Frequency
Low	Part A	<i>I'm pretty much completely lost in this area. The more time we spend on it, the more confused I get. I'm not even really sure what the "randomization test" is.</i>	52	1
	Part B	<i>predicting something?</i>		
	Part C	<i>I have no idea. . . .how close it is to that first number we found?</i>		
	Part D	<i>?</i>		
Medium -	Part A	<i>I have learned how to compare differences of subjects in an experiment. Really confused about the randomization and feel that teaching students this will cause extreme confusion</i>	96	11
	Part B	<i>Running experiments</i>		
	Part C	<i>See if we should accept/decline truths about experiments</i>		
	Part D	<i>Where can we fit this into our curriculum?</i>		
Medium	Part A	<i>The Test shows that if data sets are mixed, you can compare to data that is seperated. The point being to see if there is a difference between seperated data & the mixed data</i>	42	23
	Part B	<i>The comparison will show if your seperated data sets are actually different or different "by chance"</i>		
	Part C	<i>It really helps you decide to accept results as "by chance" or . . .</i>		
	Part D	<i>No questions, but I'm still figuring out if the confidences % (p) shows if the sets (vehicles of delivery) are diff. or same</i>		
Medium +	Part A	<i>I learned that it could do a large amount of trials in a short period of time</i>	97	9
	Part B	<i>It is useful for collecting a large amount of data using my original data (from 2 separate sets of data) and combining them randomly.</i>		
	Part C	<i>It helps me figure out if the observed value from my experiment really had a difference or if the difference happened just by chance.</i>		
	Part D	<i>How do I use it effectively with students?</i>		

Table 55—Continued

Categorization Level		Representative Response at Level	Code #	Frequency
High	Part A	<i>Initially, we found the difference of the means of 2 sets of data. Then we took the 2 sets of data, merged the data into one, shuffled the data, redistributed into 2 sets of data & computed the difference between the 2 new sets of data. We continued doing this and graphed each sets difference of means. As the # of differences increased the graph became more normal looking.</i>	07	12
	Part B	<i>It is useful to see how the initial difference compares to the graph of the many randomized differences. We were able to see how significantly different the initial sets difference compared to the many.</i>		
	Part C	<i>Helps to see if the initial 2 sets of data produced any significant difference, as in the case of determining the better vehicle in the "orbital express" experiment.</i>		
	Part D	<i>I think I'm good</i>		
				56

Several recurring issues seen in the written reflections were the mention of the words “valid” and “reliable” and some confusion about the difference between taking samples and simulating trials. Additionally, there seemed to be inconsistent use of “*p*-value” which appeared to be synonymous with alpha-level. Several teachers clung to the language of “prove” or “disprove.”

When asked on Day 4 about any questions remaining with respect to the randomization test, only 17 teachers responded with comments beyond things like “better now” or “I think I’m good.” Teachers’ responses to this prompt are listed in Table 56.

Responses are sorted according to the code assigned to reflection prompt 2 parts A - D. The majority of the responses involve either a need for more practice or logistical-type questions. Teachers 52 and 79 indicate a lack of understanding or lack of confidence relative to randomization testing. Teacher 40 indicates some confusion regarding

randomization testing and generating a sampling distribution of sample means. Teacher 71's question likely refers to trying to determine whether two measures, perhaps mean and median, are the same for a given distribution, although the intent is not entirely clear.

Table 56

Reflection 5: All Responses to Part D

Response	From R2 A - D	Code #
Still not completely positive on how to read the results & make conclusions about it.	L	52
I have not worked with it at a sufficient level of competence to have questions that are of any merit at this time.	M-	79
The only problem we continue to have is how to take more measures in Fathom. We have forgotten a step <u>every</u> time in the last few days. Today, we did correct ourselves though.	M-	68
What grade level is this idea of p-test expected to be understood with confidence?	M	02
Need practice to digest all this info on a more permanent level.	M	10
My questions would consist of me trying to keep everything straight in my mind.	M	26
How it is used in product testing. Is this the method they use?	M	37
Randomization test is pretty clear. There is a lot of new vocabulary that I'll need to use more frequently before I'll feel comfortable with it.	M	50
I think with practice at home I will be okay. Practice makes my comprehension better.	M	70
I think I need to review the meaning of this as it relates to significant and significant different.	M	2416
How do I introduce the use of technology & randomization tests w/o using a week of class time?	M	08
My questions are more logistical (where to find things) than conceptual. I need to practice on it.	M	30
Where are available lessons located online or elsewhere which could be used in a classroom to practice?	M	2371
I just need a bit more practice to solidify everything. Having access to Fathom at home may help this.	M+	57
Just need more practice to gain confidence doing this with Fathom but good with concept.	M+	66
Sometimes we replaced the value and sometimes we did not (orbital project vs. last project). Why?	M+	40
What if you want to compare 2 measures to see if they are the same?	H	71

When the reflection responses from 2 A – D and 5D are reviewed in light of the post-assessment results from Task 8 (Item 6), it is likely that teachers' understanding of

randomization testing evolved toward successful reasoning and interpreting experimental results. This conjecture is further supported with the increasing numerical ratings over the four-day professional development program representing teachers' understanding of the randomization test from reflections R2-F, R3-G, R4-A, and R5-E which are presented in the Reflection 5 section.

R2-E. Prompt R2-E was included to assess teachers' sense of the use of *CPMP-Tools* to conduct the randomization test. Fifty-three teachers responded to the prompt. Ratings ranged from 3 to 10 with mean 8.32, median 9. Figure 33 displays both numerical and graphical forms of the distribution of the ratings.

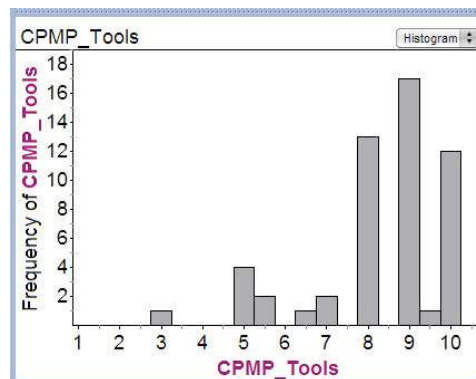


Figure 33. R2-E: Distribution of teachers' ratings of the ease of use of *CPMP-Tools*.

These ratings suggest that teachers viewed *CPMP-Tools* as relatively straightforward and easy to use, especially since they had very little time to experience the tool directly.

Reflection 3

Reflection 3 occurred following the completion of Activity 2.5 and contained the following writing prompts:

- R3-A Describe insights or new ideas you have gained through the investigations so far regarding measures of center
- R3-B Describe insights or new ideas you have gained through the investigations so far regarding variability
- R3-C Describe insights or new ideas you have gained through the investigations so far regarding bias
- R3-D Describe insights or new ideas you have gained through the investigations so far regarding design of experiments
- R3-E What questions do you have about anything we have been doing so far?
- R3-F On a scale from 1 (low) to 10 (high), rate your current feelings about how your learning of *Fathom2* is going
- R3-G On a scale from 1 (low) to 10 (high), rate your current understanding of the randomization test.

Parts R3-A – R3-E are analyzed next. Discussion of R3-F and R3-G are included with the Reflection 5 discussion along with the other prompts of these types. For each prompt R3-A through R3-D, qualitative responses were first coded with –, 0, or + to reflect the participants’ reflections as negative, neutral, or positive. Subsequently, each response was coded according to the topic(s) which it addressed.

R3-A

This prompt was designed to assess teachers’ growth in understanding of measures of center. Table 57 presents the initial categorization by professional development site.

Table 57

Reflection 3A: Nature of Teachers’ Responses by Site

	Site 1	Site 2	Site 3	Total
–	2	0	0	2
0	5	0	1	6
+	24	9	15	48
Total	31	9	16	56

The eight negative or neutral responses included three left blank and three indicating uncertainty about how to respond or confusion regarding what was meant by measure of center. The remaining 48 responses were coded to suggest something positively identified by teachers to indicate an improvement in their understanding of measures of center. The frequency of the various teacher responses are presented in Table 58. The number of responses per teacher ranged from 0 to 2; the total number of responses equals 56 in this case but does not represent one response per teacher. As mentioned above, some teachers left this prompt blank while others may have indicated multiple ideas.

Table 58

Reflection 3A: Frequency of Teacher-Reported Improved Understanding of Measures of Center

Improved understanding of statistical idea	Frequency
Mean-median relationship (skewed, outliers)	25
Randomization test (comparing centers)	5
MAD (mean absolute deviation from median)	4
Software (dynamic comparisons)	4
Teaching ideas	4
Multiple measures	3
Graphical comparisons (multiple representations)	2
Need for more than just measures of center to describe a distribution	2
Confused (uncertain)	2
Normal curve (suggesting center was top of bell curve)	2
Comparisons (generic—suggesting numerical comparisons)	1
Off target (answered different question)	1
Sample (was referring to a sample—suggesting larger sample)	1
	56

Teachers' responses suggested that many improved their understanding of the relationship between the mean and the median, particularly for skewed distributions or distributions containing outliers. The following teacher's response captures the essence of many of the responses provided:

I now understand the difference between mean & median (well, I knew the difference but I did not have a clue on how they were influenced by the data). (T2061)

R3-B

This prompt was designed to assess teachers' growth in understanding of variability. Table 59 presents the initial categorization by professional development site.

Table 59

Reflection 3B: Nature of Teachers' Responses by Site

	Site 1	Site 2	Site 3	Total
–	1	0	0	1
0	7	1	4	12
+	23	8	12	43
Total	31	9	16	56

The 13 negative or neutral responses included five left blank. The remaining 43 responses were coded to suggest something positively identified by teachers to indicate an improvement in their understanding of variability. The frequency of the various teacher responses are presented in Table 60.

The variety of responses suggests that individual teachers were gaining insights into variability in potentially different ways. These differences may have been a function of their prior experience, but they may also be a product of the numerous ways in which variability was present and/or investigated during the professional development program. Taken as a whole, the list provides a range of ways in which understanding variability may have been supported during professional development, with some aspects more or less relevant than others to individuals.

Table 60

Reflection 3B: Frequency of Teacher-Reported Improved Understanding of Variability

Improved understanding of statistical idea	Frequency
Less is better	6
Range	6
Cluster	4
Spread	4
Precision	4
Sample size	3
Relationship between measure of center and variability	3
Multiple measures (<i>SD</i> not always best)	2
Relative measures of variability	2
Quartiles	2
Standard deviation	2
Variance	2
Population vs. sample standard deviation	2
Tables & Graphs	2
Software	2
Change	1
Context	1
Definition	1
Perception	1
Role (in calculations)	1
Teaching (ideas)	1
Refreshed	2
Validity	1
Uncertain	4
Off target	3
	63

The following teacher's response captures the essence of many of the responses provided:

The less, the better—for accuracy. Understanding of this is improving each session. (T3017)

R3-C

This prompt was designed to assess teachers' growth in understanding of bias.

Table 61 presents the initial categorization by professional development site.

Table 61

Reflection 3C: Nature of Teachers' Responses by Site

	Site 1	Site 2	Site 3	Total
–	0	0	0	0
0	7	1	1	9
+	24	8	15	47
Total	31	9	16	56

Forty-seven responses were coded to suggest something positively identified by teachers to indicate an improvement in their understanding of variability. The frequency of the various teacher responses are presented in Table 62.

Table 62

Reflection 3C: Frequency of Teacher-Reported Improved Understanding of Bias

Improved understanding of statistical idea	Frequency
Human factor	14
Less is better	10
Random	8
Results	6
Rectangle activity	5
Blind/double blind	4
Teaching & assessment	3
Experimental design	2
Many types of bias	2
Definition	1
Gorilla activity	1
How and why estimators may be biased	1
Lurking variables	1
Sampling	1
Software	1
Confused/off target	3
	63

Three teachers did not respond to this prompt. The range of responses to this prompt suggest that four activities impacted teachers understanding of bias. The random rectangle activity (Act. 2.4) was mentioned directly or eluded to via the human factor and the need for random sampling. Similarly, the gorilla activity (Act. 2.4), though mentioned only once, may have impacted the human behavior reaction because of its surprisingly

deceptive expectancy reaction. The Core-Plus Mathematics Course 3 Unit 1: *Reasoning and Proof* investigation (Act. 2.5) on issues of experimental design introduced the concepts of single- and double-blind studies, and lurking variables to teachers, which were largely new concepts as judged by teachers' reactions during professional development. Finally, Standard Deviation and Its Interpretation (Act. 1.4) provided the initial discussion of the need for unbiased estimators.

R3-D

This prompt was designed to assess teachers' growth in understanding of experimental design. Table 63 presents the initial categorization by professional development site. Fifty responses were coded to suggest something positively identified by teachers to indicate an improvement in their understanding of design of experiment. The frequency of the various teacher responses are presented in Table 64.

Table 63

Reflection 3D: Nature of Teachers' Responses by Site

	Site 1	Site 2	Site 3	Total
-	0	0	0	0
0	3	2	1	6
+	28	7	15	50
Total	31	9	16	56

Two teachers did not respond to this prompt. The responses from the remaining 54 teachers support strongly that the three rules of experimental design advocated in Core-Plus Mathematics Course 3 Unit 1: *Reasoning and Proof* investigation (Act. 2.5) impacted teachers' understanding. Thirty-four teachers actually listed, described, or referred to the rules, two teachers remembered 2 of the 3 rules, and 11 of the teachers

Table 64

Reflection 3D: Frequency of Teacher-Reported Improved Understanding of Design of Experiments

Improved understanding of statistical idea	Frequency
Three rules of good experimental design (random assignment, control group, sufficient number of subjects)	34
General ideas (suggest they are more aware)	11
Control for lurking variables	7
Double blind/blind	3
Two rules	2
Teaching (having students design experiments)	2
Software	1
	60

indicated an increased awareness for the need for quality experimental design. Thus, 47 comments indicated a greater awareness of issues of three principles of experimental design and four teachers mentioned that they were already aware of these issues. Ten comments referred to lurking variables and blind or double-blind studies which many teachers communicated no familiarity with during the professional development program.

R3-E

To determine areas of struggle for teachers, this prompt asked teachers to report any questions they had about what they were learning in the session. Their responses were reviewed and then categorized by the nature of the response as shown in Table 65. Half of the teachers indicated they had no questions at the time of the prompt. Six teachers indicated a desire to slow the pace down (two teachers from Site 1) or the need to have more practice with the software (none from Site 3) and did indicate a question of understanding material. Similarly, one teacher asked a specific question about the “correct answer” to the number of ball tosses that occurred during a video segment that

had been used during the session (teachers tried to count during the video and then the distribution of their collective counts was analyzed).

Table 65

Reflection 3E: Frequency of Teacher-Reported Questions of Understanding

Categories of questions	Frequency
None	28
Teaching-related Curriculum (when & to whom) (7) Instruction (how & background to teach stats) (4)	11
Statistics-related (subcategories total more than 10 because some teachers indicate questions in 2 categories) Randomization test (4) p-value (2) variance (1) significance (1) sample size (1) lurking variable (1) non-normal distributions (1) big idea (1) probability (1)	10
Technology-related	4
Pace	2
Specific activity (gorilla video)	1
	56

Of the 11 teaching-related responses, seven were associated with issues of curriculum such as: (1) In what grades/courses should these ideas be taught? (2) At what level should this material be taught? (3) How can we put these ideas into our curriculum? and (4) What do the Grade Level Content Expectations say? For example,

I know that the MI content expectations contain much more statistics than our school teaches. Is all of this randomized testing part of that? If not, it has still been valuable for me because I've learned a whole lot about data in general & graphs & how to make sense of data [coded as teaching/curriculum]. (T2061)

Four teachers indicated some discomfort with the idea of implementing statistics in their current courses with their current statistical understanding. One Site 3 teacher's response potentially captures the general essence of the responses.

What is the expectation of math teachers teaching statistics vs. teaching what statistics is and what it is capable of & used for. After reading chapters 7 & 22, 24 in the new book you gave us, I now realize that I was thinking of stats as a branch of math rather than as another discipline. After only having one stat class I now know why I feel uncomfortable in trying to teach. I wouldn't want to teach Biology after taking only one class [coded as teaching-related/instruction]. (T2371)

Of the 10 statistics-related responses (2 from Site 1, 0 from Site 2, 8 from Site 3), a variety of responses were recorded; three of the teachers' responses are below to provide a sense of the issues.

What is a "lurking" variable--never heard this terminology; when is a sample too big (if ever) compared to the population size [coded as statistics-related/lurking variable & sample size]. (T40)

I'm still not sure about the whole "big picture." I understand how to do the whole entire problem, but when we get the "answer," or the distributions to look at, what does it mean?? What are we trying to conclude and how can we tell [coded as statistics-related/randomization test]? (T52)

Why does my wife keep buying lottery tickets when we call 1 out of 20 unlikely [coded as statistics-related/probability]? (T2693)

Reflection 4

Reflection 4 occurred at the end of Day 3, while still working on Activity 3.5, and contained the following writing prompts:

- R4-A On a scale from 1 to 10, rate your current understanding of the randomization test.
- R4-B What does "p-value" mean to you?
- R4-C What does "significantly different" mean to you?

Prompt R-4 B is discussed next, while the discussion of R4-A and R4-C will occur in the Reflection 5 section with discussion of similar prompts.

R4-B

These reflection responses were categorized representing low, medium, or high level of understanding of p -value. Table 66 provides a sample of four responses for each category. There are signs of understanding evident within each of the levels; however, the use of language and overall completeness generally increases as category increases. During the professional development program, the use of the language of “ p -value” was limited to the context of randomization testing. Forty-five of 56 (80%) teachers communicated a medium to high level of understanding of p -value at the time of this reflection prompt. It is hypothesized that the use of the randomization test was influential in teachers’ developing understanding of p -value and especially impacting their use of language referring to p -value as the probability of an unusual or “weird” experimental result occurring by chance. Though teachers’ articulation of this idea varied widely, when viewed together with their developing ideas of significant differences (see Reflection 5 section) and the likelihood of their responding correctly to post-assessment Task 8 (Item 6), it becomes more evident that teachers’ understanding of these historically challenging concepts improved.

Reflection 5

Reflection 5 occurred following the completion of Activity 4.5 and contained the following writing prompts:

- R5-A Describe what you think is meant by the phrase “the difference is statistically significant.”
- R5-B When you think about comparing distributions, how have your ideas changed or grown this week—please describe.

Table 66

Reflection 4B: Frequency and Samples of Responses by Level

Categorization Level	Representative Response at Level	Code #	Frequency
Low	<i>The p-value is the value that is desired from a randomized testing that best represents the population. The more samples taken from the population, the more they will fall closer to the mean.</i>	62	11
	<i>The p-value is the chances that an event you do not want to occur occurs! (one that disagrees with your hypothesis)</i>	61	
	<i>probability that if the sampling was to reoccur that the original data could be duplicated</i>	2416	
	<i>How confident you are about your decision</i>	23	
Medium	<i>Can tell how confident you are that the experiment will produce the results you want==>low p value mean more confidence</i>	37	23
	<i>ex: if p-value is 0.05--this means that 5% of the time, we get a value (data, result) that was unexpected (therefore due to chance) and 95% we get the anticipated result. --smaller (in general) means we have better results or that our data supports our hypothesis better than if p-value was large</i>	51	
	<i>the decimal representation of the % of data that does not fit in the norm, data that would indicate a test treatment would be a cause.</i>	82	
	<i>the lower the p-value the more likely I am to accept the results of my trials to be valid (Assuming test is OK)</i>	79	
High	<i>the value that we look at to tell if we accept or reject our hypothesis</i>	22	22
	<i>Level of confidence that an actual difference in the populations actually exists. We can never be certain or prove anything, but we can determine a level of significance. As the p-value becomes smaller the chance that the difference is by chance is smaller.</i>	10	
	<i>percentage of data beyond the original difference in mean/median graphed on the difference graph</i>	26	
	<i>p value can determine how certain you are that an event (or events) that you are testing has a result that was caused by the variable that you introduced into the testing. The smaller the p-value, the more the event is a result of your variable introduction, and not merely a random occurrence.</i>	78	
Total			56

R5-C Describe your understanding of the relationship between the size of a sample and the related sampling distribution that can be generated.

R5-D What questions do you still have about the randomization test?

- R5-E On a scale from 1 (low) to 10 (high), rate your current understanding of the randomization test.
- R5-F On a scale from 1 (low) to 10 (high), rate your current understanding of the relationship(s) between the size of a sample and the sampling distribution of the mean that can be generated from samples.
- R5-G On a scale from 1 (low) to 10 (high), rate your current feelings about how your learning of *Fathom2* is going

Responses from several other prompts from previous reflections are also reported in this section in order to compare change on similar or identical items over time.

R5-A and R4-C

These two prompts were asked during two consecutive reflection experiences with the aim of trying to determine whether and in what ways teachers' understanding of "statistically significant" may change over a short period of time (less than one day). Initially, teachers' responses to both prompts were matched side-by-side and coded in two ways: (1) the pair of responses was coded with +, 0, or –, depending on whether the combination of responses generally showed an increase in understanding (+), seemed to stay about the same (0), or appeared to decrease (–) from reflection R4-C to R5-A, and (2) the pair of responses was coded to assess the level of overall understanding of "significantly different," using high (H), medium (M), and low (L) codes. Underlying this coding scheme was the assumption that the responses would likely either improve or stay the same across prompts.

During the coding process, it became apparent that some teachers' use of language changed or the amount of writing differed dramatically from the first to the second prompt (see examples in Table 67), and it was difficult to know whether the change was actually a cognitive response or perhaps a fatigue issue. It is unlikely that the

professional development activities between prompts would have caused a negative cognitive response, however there were several examples of responses that appeared to digress. Because the responses to these prompts were particularly challenging to interpret analytically, a second approach to analysis was used.

Table 67

R4-C and R5-A: Sample Teacher Responses to Reflection Questions about “Significant Differences”

Teacher	R4-C Response	Score 1	R5-A Response	Score 2
T03	<i>That the probability of a certain event is quite low. In other words, does it occur just by chance or is it occurring way too often & not by chance.</i>	2	<i>The smaller the amount of data beyond a certain point shows an event can happen just by chance alone & be statistically significant. The more allows for not happening by chance & is not different.</i>	2
T12	<i>When comparing a random test to another, if the test is true or there is a difference do to a factor there will be significance difference between the mean & the s.d.</i>	2	<i>That there is a difference in your samples due to a factor & not just by chance</i>	3
T16	<i>A difference that would give a low p-value!</i>	2	<i>The difference is not attributable to simply chance occurrences</i>	3
T40	<i>I believe this is somewhat context bound but I generally view a p-value less than .05 as indicative of a significant difference (when viewed in the hypothesis testing format... I suppose that would be a 95% confidence interval with 2.5% in each tail)</i>	3	<i>It depends! I still like the 5% p-value</i>	2
T62	<i>Significantly different is 5% at either end of a normal distribution. This sample is possible but not very probable.</i>	2	<i>If the difference is statistically significant, the characteristic you are sampling will be evident a certain percentage of the time</i>	1
T61	<i>Two things are significantly different if there is an unreasonable p-value between them (too little similarity/correlation between the events)</i>	1	<i>The p-value is too big! (usually great than 5%). This means that the event you don't want to happen has a higher likelihood of occurring!</i>	1
T45	<i>If the "p-value" is low enough (5% example) than you can be reasonably sure that the alternate hypothesis is true because it is significant different from the sample being equally likely (null hypothesis)</i>	3	<i>That the p-value of the randomization test is small enough that you are convinced that it is not likely that this could have happened by chance</i>	3

With all of the responses in an electronic database, recoding and analysis was feasible. As a check for reliability, previous codes were removed from view, the responses shuffled to eliminate potential bias, and each individual rather than paired response coded according by articulated evidence of understanding. This time numerical codes of 1 for (low), 2 (medium), and 3 (high) were used in order to utilize mathematical operations on the results. Once the coding was complete, two comparisons of teacher's responses were made: (1) calculated *maximum*(score1, score2) to establish the highest level of understanding displayed per teacher, and (2) calculated *difference*(score2 – score1) to determine whether any progress in understanding was communicated. These results were compared to those utilizing the original coding scheme to locate and then resolve any inconsistencies.

A careful examination of teachers' responses demonstrates how challenging this concept was for many to express. Table 67 contains a sample of teachers' responses and scores for comparison purposes. Some teachers' responses suggested improved understanding (17 of 56), some stayed about the same (29 of 56) and some appeared to decline (10 of 56). Of the 10 teachers with declining scores, two were blank on R5-A.

This particular analysis was a powerful reminder of the extreme difficulty of assessing the understanding of others. For example, Teacher 40 does not likely understand less at the time of the second reflection prompt than at the first reflection prompt, but based solely on the responses provided, the demonstrated level of understanding on the second prompt is not as high as that from the first prompt. For a variety of reasons, there exists inherent variability associated with written responses, regardless of an individual's theoretical conceptual understanding.

The results of the analysis in Figure 34 indicate that 53 of 56 (95%) teachers communicated a medium to strong understanding of the meaning of “significant difference” on either the first or second prompt. This means teachers were able to articulate that a difference as large (or small) as the experimental difference was not likely due to chance or that a p -value of 0.05 or less was recorded, or a combination of both. Eight teachers’ responses on the first prompt and six teachers on the second prompt were coded as low understanding, but when both prompts were considered, only three of 56 (5%) were coded low on both.

		Score2			Row Summary
		1	2	3	
Score1	1	3	2	3	8
	2	2	12	12	26
	3	1	5	14	20
Column Summary		6	19	29	54

*n=54 because two teachers did not respond to Reflection Prompt 2 and were excluded from this summary

		Level		
		L	3	
		M	17	
		H	36	
Column Summary			56	

Figure 34. R5-A and R4-C: Summary of reflection scores and levels for “significant difference.”

The results on these two reflection prompts provide a potentially stark contrast when compared to the initial interviews conducted for this study. Of the nine teachers interviewed prior to the professional development program, only one could articulate what would be considered a medium to high level of understanding of significant difference and he was the only Advanced Placement statistics teacher in the group of interviewees. During the final interviews, all nine teachers could articulate a medium to high level of understanding of significant difference. It is likely that the professional

development program positively influenced teachers' understanding of significant difference.

R5-B

Table 68 contains the actual language provided by the teachers to this prompt representing teachers' self-reported change in understanding of *comparing distributions*. Each response was coded as positive (+), negative (–), or neutral (0). For each teacher, pre- and post-belief/comfort-level aggregate scores and content pre- and post-assessment scores are provided. The data below are sorted in ascending order according to content post-assessment score. As reported in more detail earlier in this chapter, pre/post belief scores ranged from 1 to 5; pre/post test scores ranged from 0 to 4.

One teacher did not respond to this prompt. One other teacher suggested that his/her understanding of comparing distributions had not changed. The remaining 54 teachers indicated, with some specific detail, that their understanding of comparing distributions had improved. Of all 56 teachers, two teachers (scores bolded in Table 68) had belief scores that declined from pre to post (T2371 and T06); however, their assessment scores increased. Two different teachers (scores also bolded in Table 68) had assessment scores that decreased by a small amount (T09 and T52), yet had belief scores that increased pre to post. The remaining 52 teachers saw increases from pre to post for both their reported comfort-level with statistical ideas (beliefs) and content knowledge (assessment). As reported earlier in this chapter, the change in belief scores and content assessment scores were both significantly greater than zero.

Table 68

Teachers Self-Reported Changes or Growth in Ideas about Comparing Distributions with Pre- and Post-Beliefs and Assessment Results

		Code #	Pre-Belief	Post-Belief	Pre-Test	Post-Test
They haven't changed that much. I understand a little more about finding a distribution, but I don't know how to compare them any better than before.	0	52	1.43	3.29	1.55	1.50
I have learned all new material for the first time. Wow it was a <u>lot!!</u>	+	79	2.00	3.21	1.10	1.70
In comparing distributions I know that the more random samples taken from a population will give a true picture of that population and the mean will increase giving a normal curve	+	62	1.29	2.57	1.20	1.85
To use more than one factor or reason, and how to test if the difference is due to chance	+	12	2.00	3.57	1.65	1.90
Better to have large samples, can never be totally sure of anything, important to have replacement (only random if every item has an equal chance of being drawn all the time.) Understanding the construction of a normal distribution and how to "read" it.	+	26	3.14	3.86	1.25	1.95
A person has to be very careful about snap judgments when comparing distributions	+	09	1.86	2.14	2.05	1.95
I'm experiencing growing pains. I feel that we spent so much time generating data & graphs last year, wasted time because it ended there. Not this year.	+	3011	1.86	2.86	1.90	2.10
They have changed. A more in depth study of distributions is needed before we can draw meaningful conclusions from data	+	30	2.43	3.86	1.60	2.15
I have a better understanding comparing graphs by looking at the means, st dev and ranges	+	22	1.43	2.29	1.35	2.20
It was easier to see what happens with graphs, rather than just the mathematics	+	68	2.71	2.86	1.35	2.20
I have learned tons this week about comparing distributions and am looking forward to practicing before sharing with my students. I took a lot in and now need to organize the info.	+	70	1.71	3.57	1.80	2.25
There is a lot more to look at, much of which is visual but others can be calculated. Differences between mu and x-bar, s and sigma, STD Deviations and their % of data within that deviation	+	2416	2.14	4.36	1.40	2.25
I first look at scales being used and whether there is overlapping when comparing data to tell if there is a significant difference in the data.	+	50	2.00	2.86	1.75	2.25
Before I only looked at central tendencies, now I have a better understanding of how the spread of the data and the standard deviation and the appearance of the histogram relates to give me info	+	19	2.43	3.29	1.90	2.35

Table 68—Continued

		Code #	Pre- Belief	Post- Belief	Pre- Test	Post- Test
Yes, I am much more confident with the terminology and understanding key ideas	+	3017	2.00	2.71	1.65	2.35
not just mean/median/mode; now std dev, p-values, z-scores and randomization tests	+	23	2.39	2.57	1.75	2.45
I now have expanded my vocabulary to bimodal and uniform. I, also, feel more confident at looking where mean & median lie in a skewed distribution	+	96	1.71	1.86	2.05	2.50
I can look at distributions and make educated guesses at measures of central tendency	+	3016	2.14	3.14	1.55	2.50
It is helpful to look at many attributes	+	37	2.00	2.43	1.90	2.50
I now understand how to compare a sample to a random distribution and decide if it is significant	+	08	1.86	3.00	1.70	2.60
I've seen how larger samples create distributions that have patterns that are easier to see	+	42	1.86	2.29	2.00	2.65
I have learned a great deal of information including <u>what</u> standard deviations are and how they can be used to compare dist. I never knew what skewed right/left meant/or how it affected the mean. Now I know	+	51	1.71	2.21	2.30	2.65
Grown. This is hard to articulate	+	65	1.86	3.43	2.35	2.65
I see the importance of looking at what type of distribution is present. I've also realized that what I thought was significant isn't	+	83	2.14	3.43	1.70	2.70
I don't do a lot of this in any class so this increased my awareness & refreshed my college stats days. Don't know if I'll use much in Alg 1	+	82	1.71	2.29	1.55	2.75
Comparing distributions depends on a couple of things: the scale, the mean or median (whichever is more productive in interpreting) & what we think is significant.	+	03	2.86	3.71	1.85	2.75
I can see the differences and see some ways that I can use to help my students see these differences	+	2180	2.43	3.14	1.60	2.85
I now know more about how sample size, shape and standard deviations, therefore I can more critically compare & conclude	+	54	1.57	1.86	2.00	2.85
It is very difficult to see a significant difference by only looking at the shapes of the distributions. It is necessary to look deeper before making a judgment	+	93	1.57	2.86	1.30	2.95
I did not recall working with the differences of the sample means before. I realize now how they lead us to find the actual mean and standard deviation and standard error.	+	76	3.29	3.57	1.55	2.95
This randomness test is BIG and I don't remember it, but I'm going to include it in my teachings in the future	+	78	2.57	3.71	2.25	2.90
I have a better understanding and more tools to help prove or disprove if a study is statistically significant	+	36	2.00	3.14	1.85	3.00

Table 68—Continued

		Code #	Pre- Belief	Post- Belief	Pre- Test	Post- Test
I have learned that the smaller the normal distribution the more confidence you can have in the data. The distribution shrinks when you increase the sample size.	+	20	2.14	3.29	1.95	3.00
[circled "grown"] Not all normally shaped. Look at mean & s.d. to describe a distribution	+	80	3.14	4.00	2.05	3.00
Much better idea of standard deviation meaning & how it affects shape. Also straightened out skew L & R	+	66	2.29	3.29	2.00	3.00
When comparing distributions accuracy is increased as you look at larger sample distributions	+	07	1.57	3.71	1.75	3.05
The measures of central tendency (mean, med) are a signal (and good start) but other factors (like variability) are very important also.	+	71	1.86	4.29	1.55	3.10
Initial samples and predictions/conclusions made from them can be unwise as p-scores can sometimes be high	+	16	1.71	1.93	1.85	3.10
I have more tools for interpreting data. For example randomization test, sample size comparisons and such	+	45	1.43	3.43	1.00	3.15
better understanding of why larger samples give better approximations of mean. --sometime mean/median act differently	+	10	2.57	3.00	1.60	3.15
I have a much better understanding of what to look for as far as similarities & differences and how to use the randomization test to determine the significance of the difference.	+	17	1.57	3.29	1.90	3.25
Greatly. Randomization test is a very useful tool.	+	24	2.57	3.29	1.95	3.25
Before this week I had never seen or used randomization test. I would have simply looked at the data and the summary statistics and made a conclusion	+	81	2.83	4.14	2.20	3.25
Connecting changing distributions to their graph make statistics come alive	+	2371	2.29	1.86	2.45	3.25
I'm lots more familiar with standard deviation and the mean-of-the-means!	+	61	2.14	3.00	2.25	3.25
I never thought to compare differences by randomizing. We'd calculate means/medians & stop there and compare	+	97	2.14	3.43	2.25	3.30
blank	0	06	2.29	2.14	2.30	3.30
Grown—Matching data to tables & plots has improved	+	02	1.71	2.71	2.30	3.30
I've gotten much better at identifying & understanding attributes of graphs--bimodal, skew, uniform, normal	+	39	2.29	2.57	1.85	3.35
I have a better idea of what standard deviation, standard error, normal curve, etc. mean and how they are all related to each other.	+	2693	3.14	3.57	3.05	3.35
I am less focused on the mechanics and more focused on visual representations	+	40	3.21	4.14	2.80	3.40

Table 68—Continued

		Code #	Pre- Belief	Post- Belief	Pre- Test	Post- Test
My ideas have grown from a chapterful of formulas in my advanced math book to me feeling as if I understand the meaning behind the formulas.	+	2061	1.57	3.57	1.40	3.45
I feel I have a better understanding of std. dev. & what it means in terms of the data. Ex. 68% of the data is w/in one std. dev. of mean, etc.	+	57	2.14	3.29	1.85	3.50
I have never actually performed a randomness test (just worked with canned graphs/data). It was nice to actually see it develop.	+	59	2.29	4.14	2.15	3.60
When comparing distributions and looking at various measures of center and spread as well as overall shape, conflicting indications may occur. So it is important to think about the context of the data.	+	21	4.57	4.71	3.16	3.62
small p value (less than .05 or .01--depending on situation), which allows rejection of null hypothesis and evidence to suggest alternative hypothesis	+	35	4.86	5.00	3.50	3.65
<i>M</i>			2.22	3.17	1.89	2.79
<i>Mdn</i>			2.14	3.29	1.85	2.88
<i>SD</i>			0.69	0.73	0.48	0.54

After reading the responses to the prompt, it seemed clear that the activities in which the teachers engaged during the professional development program affected their perception of their understanding of *comparing distributions*. The professional development clearly affected different individuals in different ways and some teachers struggled to articulate their thinking about *comparing distributions*, but growth in understanding of shape, center, and spread is evident. The randomization test and its interpretation were mentioned multiple times. Connections to sampling and sampling distributions were mentioned. The language of hypothesis testing was represented. Multiple representations and use of tools appeared important for some. The consistency across teachers' reports of increased understanding and reported comfort-level with statistical ideas in conjunction with teachers' performances on the content assessment

strongly suggest that the growth in their understanding of *comparing distributions* is more than random variation or regression to the mean.

R5-C and R5-F

These two prompts were designed to assess teachers' understanding of the relationship between sample size and sampling distribution. When teachers' responses to prompt R5-C were examined, they were categorized as 0, 1, 2, or 3. The categorization represented the number of aspects of the Central Limit Theorem referred to in the response with a score of 3 indicating the teacher addressed what happens to shape, center, and spread of the sampling distribution as sample size increases. Table 69 presents typical responses at each level as well as the frequency at the recorded level.

Table 69

Reflection 5: Categorization of Responses for Part C

Categorization Level	Representative Response at Level	Frequency
0	<i>Basically--reinforce many concepts and topics I already know (T35).</i>	3
1	<i>I understand that the greater the size, the smaller the standard deviation and range of means (T3017).</i>	28
2	<i>As sample size increases [up arrow], the sampling dist stand dev decreases [down arrow] and range decreases [down arrow]. The graph of the dist becomes narrower & taller. The results closer match true pop parameters as size increases [up arrow] (T39).</i>	19
3	<i>Bigger is better! Larger sample sizes produce sampling distributions that are normal and centered around mu and have a stand dev. of sigma/square root n (T71).</i>	6

When teachers rated their level of understanding of the relationship(s) between the size of a sample and the sampling distribution of the mean that can be generated from samples on R5-F, responses ranged from 3 to 10 with mean 7.71, median 8 as shown in Figure 35.

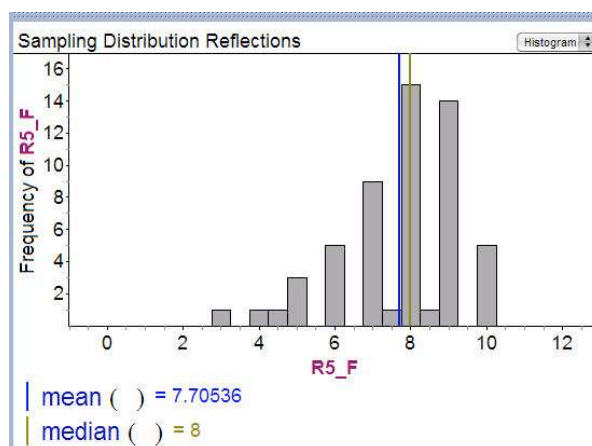


Figure 35. R5-F: Distribution of teachers' ratings of understanding the relationship between sample size and sampling distributions.

Of the six teachers with a Level 3 strong written response to R5-C, none rated their understanding on Reflection R5-F as 10. What is even more peculiar is only three of these six teachers scored at Level 4 on content assessment Item 10. That is, even though they could articulate the relationship involved in the Central Limit Theorem, they did not demonstrate an ability to consistently apply the understanding to a related set of problems. In order to further investigate these inconsistencies, additional analyses were undertaken.

Figure 36 provides a graphical representation which illustrates that teachers' self-reported understanding of the relationship between the size of a sample and the sampling distribution of sample means and their post-assessment scores on Item 10 are not strongly related.

Generally, the median reflection rating trended up with post-assessment score; however, the great deal of variability within each of the Item 10 Levels 0 – 4 distributions suggests that teachers may have thought they understood more than they could demonstrate on the post-assessment. Interestingly, there were teachers with self-reported ratings of 9 or 10 who scored 0, 1, 2, 3, and 4 on the post-assessment; some teachers with

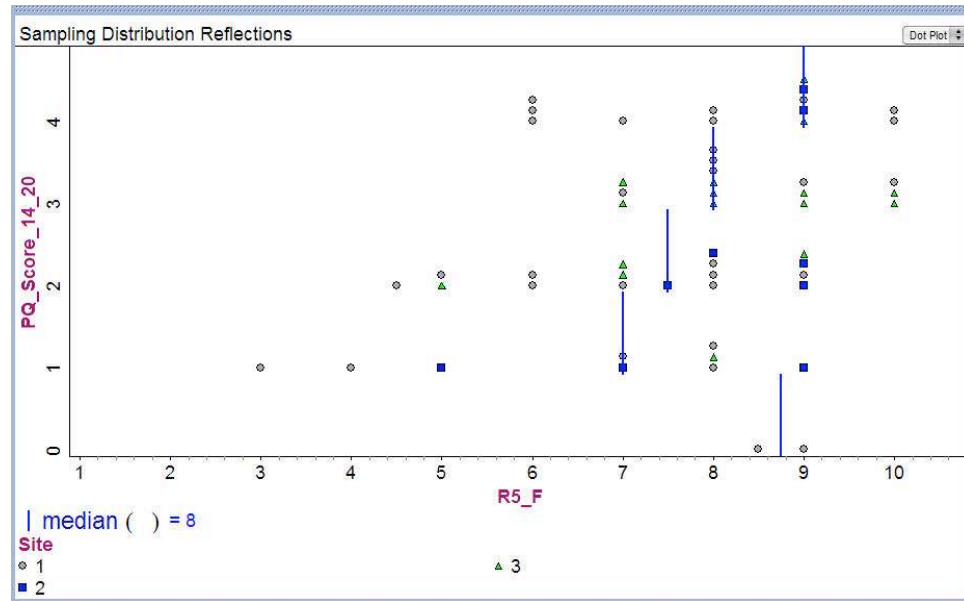


Figure 36. R5-F: Post-assessment score for Item 10 (Questions 14 – 20) compared to teachers’ self-assessed understanding of sampling distributions from Reflection 5-F.

ratings as low as 6 scored 4 on the post-assessment. Curiously, two teachers from Site 1 rated their understanding at 8.5 and 9 and then scored 0 on the post-assessment Item 10. It is entirely possible that some of the lack of strong correlation between teachers rating and their scores may have to do with fatigue at the end of an intense professional development experience. During the post-interviews, there was some mention of the fatigue issue, for example, when referring to the last page of the post-assessment and his responses, Cameron said,

This . . . this last page I wanted to let you know, I was in . . . maybe this is just a prideful thing, but I was brain dead in those last pages, so . . . Man, I was just a mush. (PostI, Cameron)

On the post-assessment Cameron chose a number of incorrect responses for Questions 14-20, but when reviewing his answers during the interview he was able to immediately and without prompting, clearly articulate correct reasoning and choices. He mentioned being embarrassed by his answers and that he should have taken more time.

This coordination of written reflections, assessment responses, and interview data converges to support the conclusion that teachers understanding of the relationship between sample size and the distribution of sample means likely improved and perhaps even more than the scores and ratings suggest.

R5-E, R2-F, R3-G, and R4-A

Teachers were asked on four separate occasions to rate their current understanding of the randomization test using a scale from 1 to 10. Table 70 contains summary statistics for each of the reflection opportunities. Teachers' ratings improved consistently over the course of the three days on which they worked with the randomization test. Furthermore, as mean ratings improved, the standard deviation of ratings for each reflection continued to decrease, suggesting both a positive shift as well as more similar ratings across teachers. It is not evident from Table 70 whether all teachers' ratings consistently improved or whether some other type of variation was happening. Table 71 contains summary statistics representing gains from R2 to R3, R3 to R4, R4 to R5, and finally R2 to R5.

Table 70

Teachers' Self-Reported Ratings for Their Understanding of the Randomization Test Across Reflection Opportunities

	R2-F	R3-G	R4-A	R5-E
n	55	56	55	56
Min	1	2	2	5
Max	10	10	10	10
<i>M</i>	6.636	7.259	7.882	8.277
<i>Mdn</i>	7	8	8	8.75
<i>SD</i>	2.111	1.758	1.650	1.198

Table 71

Change in Teachers' Self-Reported Rating for Understanding of the Randomization Test Across Reflection Opportunities

	R2 to R3	R3 to R4	R4 to R5	R2 to R5
n	55	55	55	55
Min	-2	-3	-2	-2
Max	6	5	6	7
<i>M</i>	0.627	0.627	0.382	1.627
<i>Mdn</i>	0	0.5	0	1
Change < 0	10	10	9	2
Change = 0	18	17	25	11
Change > 0	27	28	21	42

Several patterns emerge from the examination of the summary statistics from Tables 70 and 71. First, teachers' self-reported understanding of the randomization test generally improves over the course of the professional development experience, with a mean change from 6.64 to 8.28, median change from 7 to 8.75, and steadily decreasing standard deviation over time, on a scale of 1 to 10. Second, the minimum understanding rating of 1 at R2 evolves to a minimum rating of 5 by R5, clearly suggesting some progress in understanding. Finally, between each consecutive pair of reflections, 9 or 10 teachers rate their understanding lower on the more recent reflection than on the one previous. This may be an indication of a sort of implementation dip in understanding, perhaps related to the cognitive demand of coordinating conceptual understanding of the randomization test and developing technological prowess.

Upon examination of the change from R2 to R5, 53 of 55 teachers have overall change ratings greater than or equal to zero and only two teachers reported a decline in understanding. Upon closer inspection, the two teachers with reported decreases in understanding began with ratings of 9 and 7 at R2 and ended at 7 and 6, respectively. Furthermore, those with no reported gain in understanding recorded initial ratings of 7

($n = 2$), 8 ($n = 1$), 8.5 ($n = 1$), 9 ($n = 5$), and 10 ($n = 2$), suggesting a fairly strong understanding of the randomization test after one initial encounter with the process.

Though comments and justifications were not asked for on this prompt, a number of teachers volunteered comments, with the vast majority occurring on R4. This may suggest that by the time of R4, teachers were beginning to feel more comfortable sharing their thoughts, perhaps because an intellectually safe environment was emerging in the professional development setting. The teachers' comments listed below are preceded by the corresponding numerical rating given by each teacher.

The only comment from R2:

4.5 *Not enough to pass on as a teacher* (T82).

The only comment from R3:

7 *Great day!* (T3011).

Comments from R4:

2 *Low* (T42).

6 *I feel comfortable w/randomization as far as how to use it. If after completing a test, if the curve is normal & centered around 0 AND if our original diff. between means is far to either end, it means that there is data supporting the fact that there is a diff between our original groups* (T51).

7 *I'm feeling better today about the randomization test* (T97).

7-8 *Understanding it better this morning--was losing focus yesterday* (T82).

8 *I think I understand pretty well what to do* (T19).

8 *Feel good on its use and how to produce it* (T23).

8 *I am not sure why it comes out as a normal distribution* (T40).

8 *Getting clearer. I now know how to start-->null hypothesis* (T54).

9 *More use to cement it in place* (T2416).

No comments on R5.

Collectively, these results suggest that teachers in this study believed that their understanding of the randomization test improved over the course of the four-day experience. Because it is unlikely that teachers' perceptions about their understanding of the randomization test would change due to another influence, a likely interpretation is that the professional development they had been experiencing was positively impacting their understanding of the randomization test.

When viewed in conjunction with the results from the content post-assessment Item 6 (Question 8), teachers in this study demonstrated a generally strong understanding of the randomization test. Furthermore, during the post-interviews, every teacher was able to articulate an understanding of the randomization test procedure. When presented with a task for which comparing distributions and using a randomization test would be appropriate, two of nine interviewees elected to use *CPMP-Tools* to conduct a randomization test, six successfully built the randomization distribution in *Fathom2* to conduct the randomization test, and one teacher avoided using the technology but verbally discussed the procedure and its interpretation. Taken together, the written reflections, the content assessment, and the interviews provide compelling evidence that teachers' understanding of *comparing distributions* through the use of the randomization test evolved positively over the course of the professional development program.

R5-G and R3-F

At two times during the professional development program, teachers were asked to rate their current feelings about their learning of *Fathom2* on a scale from 1 (low) to 10 (high). Table 72 contains summary statistics for each of the reflection opportunities.

Table 72

Teachers' Self-Reported Ratings for Their Feelings about Their Learning of Fathom2

	R3-F	R5-G	Gain
n	56	56	56
Min	2	5	-3
Max	10	10	6
<i>M</i>	6.679	7.304	0.625
<i>Mdn</i>	6.75	7	0.5
<i>SD</i>	1.833	1.448	1.573

The mean gain score of 0.625 is significantly greater than 0, with $p = 0.002$, $t_{55} = 2.973$ using a paired t -test. Given the sophisticated way in which teachers were encouraged to use *Fathom2* as a tool, the modest gain in ratings may suggest that they were not especially confident in their skills. It is encouraging that teachers became relatively comfortable using *Fathom2* in such a short period of time. As they created the mechanisms to simulate randomization distributions and sampling distributions of the mean in order to conduct investigations, teachers needed to simultaneously coordinate new statistical understanding and facility with new technology. The cognitive demands on learners through the conduct of these investigations were extensive.

Summary of Reflections

Written reflections occurred at five distinct times over the course of the four-day professional development program. In all, 24 reflection prompts were provided for 56 teachers and their responses analyzed. Each reflection prompt provided formative assessment data for the facilitator of the sessions during the professional development as well as data for the retrospective analysis associated with the study. Without exception, responses to each prompt appeared to indicate that teachers believed they were coming to understand and having new insights about statistical concepts associated with *comparing*

distributions. Teacher participants in this study were enormously generous with their written communication during reflections, which both informed the study while creating a mountain of qualitative data for analysis. While traversing through the signals from the 24 individual reflection items en route to the summit of *Teacher Understanding*, teachers' evolving understanding of *comparing distributions* continued to emerge. Teachers' improved understanding appears to evolve from initial new ideas about measures of center and spread as seen in R1, to beginning use of dynamic statistics software and initial conceptions of the randomization test in R2, to design of experiments and use of dynamic statistical software to conduct the randomization test in R3, to issues of hypothesis testing in R4, and finally to general growth and understanding across multiple ideas including the relationship between sample size and sampling distributions in R5. This trajectory, in conjunction with prior analyses will form the basis of the final chapter.

Transitioning to Chapter VI

Thus far, the presentation of data and analyses have illuminated classroom interactions in the context of statistical professional development for high school mathematics teachers, changes in teachers' statistical content knowledge and beliefs about that knowledge, reactions to activities through reflecting on those activities, and voices of teachers sharing their understanding of *comparing distributions*. The next chapter, Chapter VI, will coordinate the results and analyses from Chapters IV and V in order to answer the research questions.

CHAPTER VI

CONCLUSION

If one takes the position that an understanding of statistics is a necessary component of education for a literate, socially responsible society, an obvious consideration is how to educate students within the constraints imposed upon schools by virtue of the pressure of high stakes accountability, current teacher preparation and professional development programs, curriculum and instructional materials, availability of technology, and myriad additional societal variables. The present study assumed that the statistical education of students requires teachers with an understanding of important statistical concepts and connections among concepts. Because of the potentially large number of high school teachers for whom statistical reasoning remains a veritable mystery, this study sought to engineer a professional development experience that might honor teachers' time, knowledge, experience, and dispositions, and support their statistical understanding in an effective and efficient way. Design research was appropriate because it is particularly useful when creating and studying the efficacy of innovative learning environments.

It was hypothesized that, through participation in a four-day professional development program through which teachers explored *comparing distributions* as described in this study, with the purposeful use of the dynamic statistical tools *CPMP-Tools* and *Fathom2* and learning activities and sequences engineered with potential to

support teachers' understanding of *comparing distributions*, teachers' understanding would improve. Teachers' understanding, as interpreted in this study, emerged in use (Bakker, Derry, & Konold, 2006). Three separate iterations of the instructional design were implemented with three groups of high school mathematics teachers at three professional development sites ($n_1 = 31$, $n_2 = 9$, $n_3 = 16$).

This chapter coordinates and synthesizes the analyses and results from Chapters IV and V in order to answer the three main research questions. A model with potential for supporting teachers' evolving understanding of *comparing distributions* will be presented and elaborated. Possible implications arising from the study will be discussed and limitations to the study presented. The chapter concludes with suggestions for future research based on experiences with this study.

Answering the Research Questions: Bracketing the Researcher's View

In contrasting and describing this research study, the words “knowledge” and “understanding” have been used interchangeably. This use of language represents the researcher's personal perspectives on teaching and learning. For the researcher, to “know,” means to understand deeply and broadly. In the colloquial sense of the words, the researcher acknowledges that one can “know” something about something and not fully “understand” all there is to know about it.

Like Boaler and Greeno (2000), the researcher considers “knowing and understanding mathematics as aspects of participation in social practices, particularly discourse practices, in which people engage in sense-making and problem-solving using

mathematical representations, concepts, and methods as resources” (p. 172). They speak of knowing as “connected knowing,” in which

the individual considers her knowledge as primarily being constructed in interaction with other people (either directly, in conversation, or indirectly, through interacting with texts or other representations of others’ knowledge and thinking), in a process that depends on understanding others’ experiences, perspectives, and reasoning, and incorporates this understanding into the individual’s knowing and understanding. (Boaler & Greeno, 2000, p. 174)

In the researcher’s view, knowing and understanding are inextricably linked and it was not the goal of this study to dissect them. Instead, evidence of knowing and understanding was viewed through the discourse practices of reasoning and justifying. Changes in knowing and understanding were seen as changes in teachers’ reasoning and justifying in the context of written text, conversations, and classroom dialogue. Bakker, Derry, and Konold (2006), arguing from the philosophical view of Brandom (2000), proposed that an inferential view of teaching and learning statistics should be privileged over a referential view. They explicated a view of understanding that resonates with the researcher:

We will characterize a *referential* view as focusing on concepts and graphs as representations or mirrors of some reality—whether physical or ideal. An *inferential* view sees grasping a concept or understanding a graph as mastering the use of the word or graph in a process of reasoning. Knowing, in the inferential view, is seen as participation in a social practice of giving and asking for reasons, and committing to the inferences that are implicit in making those claims. Participation does not require an immediate and full grasp of the explicit meaning of reasons and claims but rather, the ability to inhabit the space in which they operate. Understanding thus emerges in use. (Bakker et al., 2006, p. 1)

Bakker et al. referred to the work of Brandom (2000) who suggested that concepts come in packages and that in order to have a concept, one must have many concepts because “Cognitively, grasp of just one concept is the sound of one hand

clapping” (Brandom, 2000, p. 49). This position is consistent with that of the National Research Council (2001): “*Conceptual understanding* refers to an integrated and functional grasp of mathematical ideas” (p. 118).

The researcher’s perspective in this study is that one can never know precisely what another person knows or the ways in which s/he know it. At best one can approximate another’s knowledge. In the tradition of the situative perspective, mathematical knowing is seen as “sustained participation in mathematical practices” (Greeno, 2003). In this study, sustained participation was limited to four days (20 hours) of a professional development experience. Accurate or believable approximations of complex phenomena, such as knowing or understanding, can be strengthened when multiple sources of data are used for the approximation. When numerous sources of data of differing types tend to converge to a signal, one can begin to have some degree of confidence in a conclusion. As the following sections will demonstrate, multiple data sources were coordinated to analyze the three research questions in this study.

Question 1: Knowledge of Comparing Distributions

The first research question was, “What do high school mathematics teachers know about *comparing distributions*?” The answer to this question will be presented in two parts. The first part will address teachers’ knowledge prior to the professional development associated with the study. The second part will address teachers’ knowledge following the completion of the professional development experience. As a means of approximating what teachers’ know about *comparing distributions*, the primary sources of data utilized were written pre- and post-assessments and interviews with a subset of

teachers. Additional ways of knowing were investigated during the professional development experience and viewed through the video artifacts and teacher written reflections. Data from all of these sources informed the answer to this question.

Teachers' Knowledge of Comparing Distributions Prior to Professional Development

Teachers' knowledge of *comparing distributions* at the beginning of the study was approximated using written pre-assessments of 56 teachers and pre-interviews with a subset of 9 teachers. These results served as baseline data from which to look for change over the course of the study. It should be noted that these data document the consequences of the teachers' prior instruction in or experiences with statistics (P. Cobb, 2000). Results were interpreted using the four-level framework provided in Table 5 in Chapter III. From the analyses of those data, the 56 teachers in this study performed in the range of 1.00 to 3.50 on a four-level scale with four representing the highest score possible ($M = 1.89$, $Mdn = 1.85$, $SD = 0.48$). The vast majority of teachers (44 of 56, 79%) performed in the neighborhood of Level 2 at the pre-assessment. An important finding here was that teachers' pre-assessment scores were independent of their prior statistics college coursework. Teachers with no formal statistics training scored at approximately the same level as those with 1, 2, or 3 statistics courses.

Figure 37 presents pre- and post-assessment distributions for teachers' overall scores and subsets disaggregated by the statistical big ideas related to *comparing distributions*: distribution, variability, and sampling distributions. These graphical representations augment the numerical results reported in Chapter V. The histograms in this figure were configured to correspond to the four-level theoretical framework in the study. For purposes of classification and interpretation, scores within a level will be those

within plus or minus one-half level from a given level. In other words, scores are rounded to the nearest level. This figure will be referred to throughout this section.

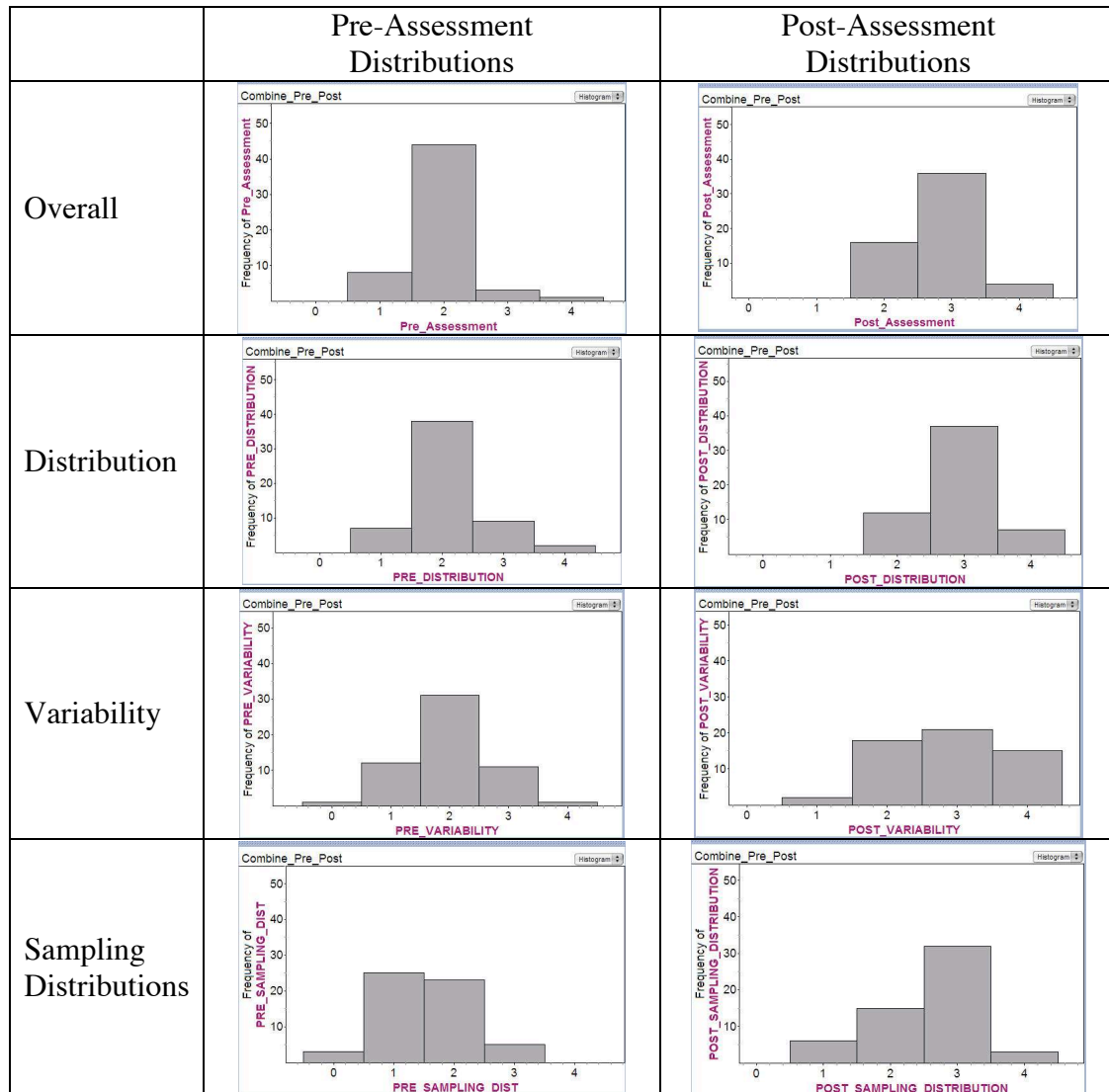


Figure 37. Comparison of pre- and post-assessment distributions overall and by big idea.

Level 2 on the pre-assessment represents the emerging distributional level of *comparing distributions*. This level represents the

first holistic view of the data . . . where informal qualitative descriptors of the data, along with basic summary statistics, are used to describe two datasets. Teachers begin to understand the difficulty in creating measurable conjectures, but are unable to successfully resolve the conflict and show frustration in attempting to write an appropriate conjecture.

Variability, while acknowledged, is not understood beyond a descriptive level. (Makar & Confrey, 2004, p. 3)

With respect to variability, teachers demonstrated partial recognition of variation (put ideas in context, tend to focus on single aspects and neglect others) (Watson et al., 2003), and for sampling distributions, they are at the verbal reasoning stage (have a verbal understanding of sampling distributions and implications of the Central Limit Theorem, but cannot apply this to the actual behavior of sample means in repeated samples) (Chance et al., 2004). Of the three big ideas associated with *comparing distributions*, as expected, sampling distributions was the most challenging area for teachers. This is consistent with the research of others (Chance et al., 2004; Makar, 2004).

Pre-interviews generally confirmed the results seen from the written pre-assessment. Teachers demonstrated a procedural knowledge of basic descriptive statistics and the ability to read graphical displays in the forms of histograms, boxplots, and dotplots. Their interpretations of the graphical representations were typically limited to comparing surface features of the graphs. With the exception of the few teachers with some experience teaching statistics, teachers were not confident in their statistical knowledge. The recollections of their college statistics experiences suggested their opportunity to learn statistics was largely theoretically- and probabilistically-based and with little or no connection to investigations or technology. Language used by teachers suggested a basic knowledge of mean, median, mode, range, IQR, minimum, and maximum values. Some teachers mentioned standard deviation but typically were unable to articulate an interpretation of its meaning. Teachers did appear to reason about distributions as objects, rather than focusing on individual data points, contrary to reasoning patterns associated with research on school-age students (Bakker &

Gravemeijer, 2004; Konold & Higgins, 2003), although there were exceptions to this rule. Many teachers *compared distributions* with some attention to center (typically seen as middle cluster, mean or median), variability (typically seen as range), and shape (typically referred to as normal or skewed), but for each, errors in reasoning were common. Teachers' sense of statistically significant difference was typically limited to either a contextual-based rationale or simply the magnitude of the difference in some measure of center (typically the estimated mean or median) was large. For the teachers in this study, the statistical process uniformly excluded any attention to the formulation of a researchable question.

As Cobb (2000) suggested, results such as these may be viewed, in part, as documenting the consequences of the teachers' prior instruction in statistics. The written assessment, in conjunction with the personal interviews, provided some idea of the content and quality of prior instruction. Interestingly, like the seventh graders in Cobb's study, the teachers in the present study likely experienced statistical classrooms where "classroom activities had emphasized calculational procedures and conventions for drawing graphs rather than the creation and manipulation of graphs to detect trends and patterns in data" (Cobb, 2000, p. 47). Nearly every teacher interviewed described their statistical coursework as focused on procedures, mechanics, and theory. As expressed during the professional development program, and reinforced through their written reflections, exploring datasets, reasoning from multiple graphical representations, and making statistical arguments were new to most teachers. Though the calculational procedures likely emphasized in collegiate statistics courses for these teachers far

exceeded the level seventh graders would experience, it may be the case that the results are parallel but perhaps with more detail or components.

Exploratory data analysis (EDA) did not appear part of teachers' statistical repertoire and there was no evidence found suggesting the use of technology for doing statistics beyond one or two-variable statistics on a TI-84 graphing calculator. Like Cobb (2000), viewing the teachers' statistical reasoning as situated with respect to prior instruction enabled careful thinking with respect to reasonable starting points for the design experiment described herein. Prior to the professional development experience as part of this study, teachers exhibited knowledge and practices of statistics that appeared inadequate to meet the demands of teaching high school students statistics in the spirit suggested by GAISE (Franklin et al., 2007) and PSSM (NCTM, 2000). They did, however, exhibit a number of productive data-based reasoning strategies that possibly formed the basis for their growth in understanding.

Teachers' Knowledge of Comparing Distributions Following Professional Development

Post-assessment data indicate that teachers' understanding of comparing distribution improved significantly from that evidenced on the pre-assessment. All four post-assessment distributions in Figure 37 exhibit positive distributional shifts from the paired pre-assessment distributions. Overall, the 56 teachers in the study performed in the range of 1.50 to 3.65 on the four-level scale ($M = 2.78$, $Mdn = 2.88$, $SD = 0.54$) on the post-assessment. The majority of teachers (36 of 56, 64%) performed at approximately Level 3 at the post-assessment. Gain-scores were significantly greater than 0 for all comparisons (overall—effect size 1.82; distribution—effect size 1.54; variability—effect

size 0.99; sampling distributions—effect size 1.26). Teachers with post-assessment scores at Level 4 began with pre-assessment scores at Levels 2, 3, or 4 (see Table 5 in Chapter III for detailed descriptions of levels). Teachers at Level 3 began with pre-assessment scores of 1, 2, or 3. Teachers at Level 2 began with pre-assessment scores of 1 or 2. There was clearly a change in teacher understanding as evidenced by differences in assessment scores pre to post. The positive distributional shifts from pre to post for the big ideas of distribution, variability, and sampling distributions are consistent with the overall positive change. Notably, the shift in improvement seen in the understanding of sampling distributions was greater than that of distribution and variability, suggesting that the professional development intervention may have positively impacted teachers' understanding of this particularly challenging statistical idea.

Level 3 on the post-assessment represents the transitional view of *comparing distributions*. At this level, teachers

begin to understand the influence of variability in comparing two groups; more flexibility is shown (e.g., multiple graphical representations, alternative measures of center or spread) in comparing datasets at this level; conjectures, while questionably measurable, have progressed to show elementary understanding of the difficulty in creating a conjecture that does not overly compromise the question at hand, but allows for possible collection of data. (Makar & Confrey, 2004, p. 3)

With respect to variability, Level 3 represents applications of variation (consolidating and using ideas in context, inconsistent in picking salient features) (Watson et al., 2003).

Understanding sampling distributions straddles two of the categories: (1) transitional reasoning (able to correctly identify one or two characteristics of the sampling process without fully integrating the characteristics), and (2) procedural reasoning (able to

correctly identify the three characteristics of the sampling process but does not fully integrate them or understand predictable long-term process) (Chance et al., 2004).

What teachers came to understand about *comparing distributions* may have been conservatively estimated with the written post-assessment. Some teachers indicated during the post-interview that, at the time of the post-assessment, they were mentally tired from the week of professional development and may not have been at their best. Based upon the analysis of the written reflections and post-interviews, teachers were able to articulate verbally and in writing, a much improved command over the language of statistics for making comparisons. As seen throughout Chapter 5, teachers believed and communicated that their understanding of *comparing distributions* was improving. Their increased facility with *CPMP-Tools* and *Fathom2* software went from no experience to surprisingly proficient in a very short period of time. Teachers' data-based arguments evolved to include *comparing distributions* with respect to multiple measures of center and spread, multiple graphical representations, attention to sample sizes and experimental design, the importance of context for interpretation, and ultimately to whether measures from two independent samples were significantly different. Teachers appeared to be growing more confident in their understanding and more skeptical about making snap decisions. With technological tools, teachers tested their ideas and challenged their conceptions.

There were still instances of teachers whose understanding of some statistical concepts, based upon written or verbal responses, would be considered fragile or lacking a strong statistical perspective (referring to Levels 1 or 2 from the framework in Table 5, Chapter III). Most notably, in the resampling environment, some teachers referred to the

process of resampling as taking more samples rather than conducting more resampling trials and they may have been confused. It is unclear whether this confusion was conceptual or semantic. Some teachers continued to associate more normal-looking distributions with less variability than non-normal-looking distributions when making comparisons. There was at least one teacher who associated the randomization test process with adding more data to the experiment rather than resampling from the existing data. One teacher referred to the “sample population,” which may have been an artifact from the session or previous coursework, but it also may have been in reference to the “sample collection” in *Fathom2*. Though these responses were not common, they are useful to inform future iterations of this work. The examples can form a basis for continuing to improve the design for subsequent experiments.

Using the GAISE (Franklin et al., 2007) framework as a guide, whereas teachers’ demonstrated level of understanding at the beginning of the study would have been difficult to characterize as consistently at Level B (loosely associated with middle school level) or above, by the end of the four days of professional development, teachers’ demonstrated understanding (excluding bivariate measurement data which was not part of this particular intervention), could easily be mapped to Level C (loosely associated with high school level). At Level C, the statistical problem-solving process components include formulate questions, collect data, analyze data, and interpret results. At this level, learners pose their own statistical questions of interest and the questions seek generalization. Learners design for differences utilizing sampling designs with random selection and experimental designs with randomization. Learners understand and use distributions in analysis as a global concept; they measure variability within and between

groups, compare group to group using displays and measures of variability, and describe and quantify sampling error. Learners are able to look beyond the data in some contexts to generalize from sample to population, they are aware of the effect of randomization on the results of experiments and understand the difference between observational studies and experiments, and they can distinguish between conclusions from association studies and experiments. Finally, learners attend to chance variability in addition to sampling variability, measurement variability, natural variability, and induced variability (Franklin et al., 2007).

The analyses and results reported in Chapters IV and V strongly support the claim that teachers in this study had the opportunity to explore all of the competencies associated with *comparing distributions* at Level C as outlined in GAISE. The same analyses provide evidence that teachers' understanding of *comparing distributions*, including distribution, variability, and sampling distributions increased significantly.

Question 2: Resampling Techniques and Dynamic Statistical Tool Usage

The second research question was, "How do professional development experiences with resampling techniques (RT) and dynamic statistical tools (DST), as described in this study, shape what teachers know about *comparing distributions*?" To answer this question, analyses relied heavily on the video of the professional development program, teachers' written reflections, and post-interviews. As seen in the previous section, teachers' understanding of *comparing distributions* changed significantly from pre- to post-assessment, thus minimally it is fair to say that experiences with resampling techniques and dynamic statistical tools supported teachers' learning

about *comparing distributions* as both were significant components of the experience. For clarification, DST will encompass the use of *CPMP-Tools* and *Fathom2* software. Later the case will be made to include a non-technological but still dynamic statistical tool, but for now the discussion will include only the two mentioned. Also, when referring to RT, the discussion will be strictly limited to the randomization test procedure (also known as permutation testing) and the generation of the sampling distribution of sample means as those were the two examples of resampling techniques that were used and investigated in the course of this professional development. Resampling, in this case is used to imply a Monte Carlo approach to simulation by resampling under appropriate conditions to generate an empirical sampling distribution from which to reason.

The discussion of the use of DST and RT assumes a classroom environment supportive of investigation, conjecturing, ongoing discourse, and interaction between facilitator and teachers and teachers and teachers. Without a learner-centered environment in which all members of the classroom community work collaboratively to build toward a shared understanding of mathematical and statistical ideas, it is unlikely that what is about to be described would actually occur. A brief chronology of the use of DST and RT in the professional development is provided and a more thorough treatment of the discussion may be found in Chapter IV.

The initial activities in the professional development program made modest technology assumptions. The introduction of technology was designed to build on, extend, and attempt to support connections to the representational competencies that were shared within the professional development classroom. Teachers were introduced to *Fathom2* through the facilitator's modeling the creation of a collection of data from the

teachers and attempting to induct them somewhat covertly into the navigation world of *Fathom2*. By verbalizing the navigational moves as she generated tabular, numerical, and graphical representations, the facilitator was planting content and pedagogical seeds. Teachers began to make comparisons almost immediately once graphical representations were available for comparing. The facilitator modeled dragging a data value from a distribution represented as a dotplot and demonstrated the effect on the mean and the median under varying conditions as teachers made conjectures. Periodically throughout the session, as investigations warranted, *Fathom2* was utilized through demonstrations of quick data gathering and representation-comparing illustrations in the context of the current investigation. Later, one custom tool of *CPMP-Tools* was used to develop conceptual understanding of mean as the balance point of a histogram and another to develop standard deviation as the distance from the mean needed to encompass two-thirds of the data from a normal distribution (Site 1 and Site 3 teachers experienced these features through demonstration only, due to technical difficulties.). *CPMP-Tools* was used to automate the randomization distribution which had been only physically simulated up to that point. Then teachers explored two (some three) guided tours of *Fathom2* in order to develop some personal capacity to navigate in the environment. Shortly thereafter, the facilitator modeled building the randomization distribution mechanism for teachers and they then built their own. In the context of additional statistical investigations, teachers gained facility with constructing the randomization test mechanism and reasoning from the resulting randomization distribution. Teachers were then introduced to function plotting in *Fathom2* through the exploration of the parameters of the normal density function using sliders. The final new capability teachers learned

was to construct the mechanism to generate the sampling distribution of sample means and revisited function plotting to explore the mathematical model relating sample size to standard error of the mean. Finally, the last activity with DST involved an investigation in which teachers would ultimately conduct a randomization test and use either *CPMP-Tools* or *Fathom2* for the simulation. During the course of activities, once *Fathom2* had been introduced, it was used frequently as questions arose.

Multiple representations, and particularly hot-linked dynamic representations, provided immediate feedback and oftentimes challenges to teachers' intuitions and predictions. As representations were easily morphed, teachers were challenged to reconcile their conceptions with those presented in other representations. The ease with which they could make and test conjectures about data or representations was likely a major mediating factor. Potential for making connections between representations and concepts was rich. *Comparing distributions* sometimes meant comparing the same distribution but with alternate representations or measures. *Comparing distributions* sometimes meant comparing two distributions with multiple measures of center or variability. *Comparing distributions* sometimes meant comparing sampling distributions resulting from varying sample size constraints or number of trials. Along the way, teachers were challenged and supported to make sense of the representations, the context, the data, and the resulting statistical argumentation that developed collaboratively among the groups. Table 73 provides a condensed view of the contributions of RT and DST to teachers' understanding of comparing distributions.

Again, it is critical to note that these tools and techniques did not manifest in a vacuum. They were situated in an environment in which teachers' learning was central.

Supports for the environment included rich, investigative tasks, a spirit of collaboration and respect, and purposeful challenge and support for learning. The characteristics of professional development in addition to the use of RT and DST are the topic of the third research question which is presented next.

Table 73

Resampling Techniques and Dynamic Statistical Technology: Contributing to Teachers' Understanding of Comparing Distributions

Who?	What?	How? Contribution to shaping teachers' understanding of comparing distributions
Facilitator	With <i>Fathom2</i> , established collections, generated tables, graphs, summary statistics, and multiple-linked graphs for comparison of distributional features. Compared measures while dragging data values	Introduced the dynamic nature of the software and helped to connect distributional measures of center and variation to distributions with dynamically changing characteristics. Confronted procedural view of mean and median with a more relational view. Provided evidence that distributions can have the same measures but qualitatively different from other features (e.g., may match mean and range, but with vastly different shapes).
Teachers	Explored <i>CPMP-Tools</i> balancing histogram and estimating standard deviation tools	Established mean as balance point and standard deviation as distance from the mean representing 68% of data of normal distributions. May have inadvertently contributed to some teachers' confusion about standard deviation in non-normally distributed samples. For others this seemed to augment their understanding of mean and standard deviation with graphical representations and connections.
Facilitator/ Teachers	Demonstrated and used <i>CPMP-Tools</i> randomization distribution tool to conduct the randomization test	Established the difference of means of two independent samples as a measure with long term predictable variability; established the linkage of repeating trials and looking for long term behavior; introduced sampling distribution concept; established p -value as the probability associated with the empirical frequency distribution. Served as a linkage from the physical simulation environment to the technology-based construction environment.
Facilitator	Demonstrated the construction of the randomization distribution mechanism in <i>Fathom2</i>	Illustrated the hierarchy of objects in the simulation (population, sample, collection of measures). Middle tier highlighted sampling variability under the null hypothesis. Provided contrast between histogram and boxplot representations with the original data compared to the resampled data. Animation feature allowed user to simultaneously view the resampling, its effects on the individual resample, and the contribution to the generation of the randomization distribution.

Table 73—Continued

Who?	What?	How? Contribution to shaping teachers' understanding of <i>comparing distributions</i>
Teachers	Constructed the randomization distribution mechanism in <i>Fathom2</i> and used it to <i>compare distributions</i>	Reinforced the hierarchical nature of the development of a sampling distribution. Made connections between population and sample explicit, though this was a place where some teachers may have failed to make this connection. Sampling variability was explored as was the behavior of the randomization distribution using different contexts for exploration.
Facilitator	Demonstrated function plotting using sliders for parameters in mathematical models	Connected the somewhat familiar mathematical work algebra and modeling with the context of statistical models. Provided visual connections to the normal density function while dynamically change mu and sigma. Compared multiple normal density curves by dynamically changing mu and sigma with sliders, connected to algebraic transformations.
Facilitator	Demonstrated the construction of the mechanism to generate the sampling distribution of sample means from a given population	Extended a physical simulation experience to a dynamic environment. Animated sampling highlighted the sampling variability of individual samples of varying sizes. Generated initial conjectures about the relationship between sample size and the sampling distribution of the mean.
Teachers	Constructed the mechanism to generate the sampling distribution of sample mean and used a self – selected or generated population to explore the behavior of the sampling distribution.	Explored the behavior of samples of varying sizes from a population of their choice. Regularities were documented, accumulated and compared. Comparisons included sample size, mean, standard deviation, range, and shape. Teachers' proficiency with generating representations and making comparisons noticeably improved.
Teachers	Used function plotting in <i>Fathom2</i> with sliders to investigate the mathematical relationship between sample size and standard error of the mean.	Provided a quantifiable relationship between sample size and standard error and a direct connection to the Central Limit Theorem. The mathematical modeling connections appeared to honor and extend teachers' prior knowledge and provide access to this powerful statistical relationship.
Facilitators/ Teachers	Used <i>CPMP-Tools</i> and <i>Fathom2</i> to explore data from an experimental research study and compare experimental and treatment conditions.	Reinforced comparing distributions with respect to numerical and graphical representations; created cognitive dissonance for teachers and challenged once again, their conceptions of significantly different, affirmed the randomization test as a viable vehicle for quantifying differences between distributions.

Based on extensive retrospective analyses, as noted earlier, some of which was reported at the end of Chapter IV (Table 17), factors or themes associated with RT and DST that emerged as potentially influential in promoting teachers' understanding of

comparing distributions were (1) making statistical arguments, (2) sense-making, (3) multiple representations, (4) dynamic statistical software, and (5) resampling. Making statistical arguments in the DST and RT environment included foci on pressing beyond “right answers,” privileging reasoning (the inferential view), encouraging divergent thinking, requiring evidence for statements when putting forth statistical arguments, connecting statistical and mathematical reasoning in appropriate ways (e.g., Activity 3.4), and articulating thinking verbally and in writing. The focus on sense-making shared the characteristics of pressing beyond “right answers,” privileging reasoning, and articulating thinking verbally and in writing. It also incorporated making predictions, coordinating representations and perspectives, and experiencing cognitive dissonance. The focus on multiple representations included moving between graphical, tabular, and symbolic representations; attending to representations in the forms of histograms, dotplots, boxplots, function plots, and summary tables; attending to shape, center, and variability; and the use of *CPMP-Tools* and *Fathom2* software to support access and availability of multiple representations.

The presence and use of *CPMP-Tools* and *Fathom2* software in the professional development were critical for providing teachers’ access to representations that would otherwise have made the investigation of resampling techniques with dynamically-linked representations logistically infeasible. Similarly, activities and tasks designed to investigate statistical phenomena from a perspective of resampling made access to dynamic technology mandatory. The presence of DST contributed to the ability, in a technological environment, to explore the randomization distribution and the sampling distribution of sample means. It provided a dynamic medium through which to explore,

contrast, and link distributional characteristics of shape, center, and variability. At times, it provided a scaffold for learners to extend their statistical reasoning and understanding as it supported their exploration of relationships within and among representations.

Resampling, as described in this study, was limited to the process of randomization testing to compare independent samples and building the sampling distribution of sample means from a given population. The use of *CPMP-Tools* and *Fathom2* allowed learners to explore the long run behavior of empirically-constructed sampling distributions and provided a scaffold for the learner to connect and extend prior statistical understanding. Perhaps most importantly, the act of resampling to ultimately generate a sampling distribution in a technological environment allowed the hierarchy of objects in the sampling process to remain visible to the learner (Lane-Getaz, 2006; Saldanha, 2004).

As a potentially useful way to envision some of the complexities involved in understanding how DST and RT helped to shape teachers' understanding of comparing distributions, a model from previous work in algebra was utilized. Van De Walle (2004) provided a view of five representations of functions as a complete digraph with five vertices. The vertices of the model were labeled (1) language, (2) context, (3) table, (4) graph, and (5) equation, and visually represented the connections among five different representations of function. With a slight modification, replacing the "equation" vertex with one labeled "measures," the resulting digraph appears to capture the relationships among the representations used when *comparing distributions* in this study (see the inner circle in Figure 38). As evidenced through the sequencing of learning activities and the enacted learning trajectory in Chapter IV, the purposeful and continual movement within

and among representations of language, context, tables, graphs, and measures, which was largely made possible through the flexible use of DST and encouraged in the context of RT, may have been particularly influential in shaping teachers understanding of *comparing distributions*. As a supplement to the original conceptual model of big ideas of *comparing distributions* through attending to relationships among distribution, variability, and sampling distributions, explicit attention to multiple representations and their contribution to supporting all three of the big ideas associated with *comparing distributions* likely improves the model (see Figure 38).

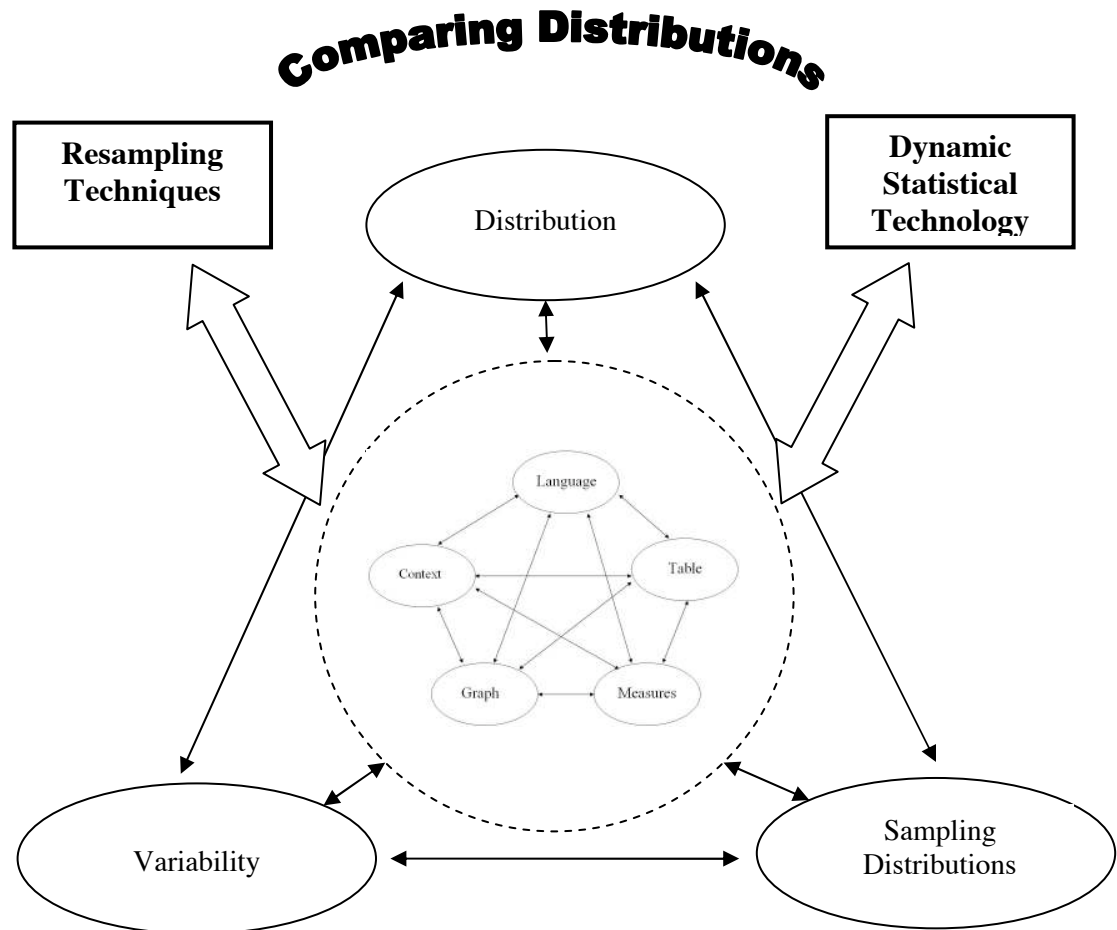


Figure 38. A model for the relationships among content, representations, and technological tools supporting teachers' understanding of comparing distributions.

Consistent with the Vygotskian view that tools serve as mediators of learning, DST and RT are seen as tools mediating learning of *comparing distributions*. In particular, DST and RT synergistically mediate learning through the presence of and access to dynamically-changing, user-controlled, multiple representations. The complex relationship among the big ideas of *comparing distributions*, multiple representations of *comparing distributions*, and meditational factors of RT and DST as described in this study may be seen in Figure 38 (Note: size of text and arrows reveal relationships only; they are not meant to imply importance). Together DST and RT may provide means and opportunity for learners to interact with their personal statistical conceptions, transform and evaluate statistical objects in order to make comparisons, reconcile discrepancies they might discover, and make connections within and between concepts. The interactions with the DST and RT may differ from learner to learner. According to Zbiek and colleagues (2007),

The construct of instrumental genesis is helpful to researchers in examining the role of technology in learning. It explains how technology does not have the same automatic power for all users and how its intelligent use requires both conceptual and technical knowledge.
(p. 1179)

By instrumental genesis, they mean the process by which an artifact becomes an instrument. In this study, one way in which the use of DST and RT was found to have differential impacts on learners was discovered during the post interviews. The first interview prompt was designed as a performance task for which the use of DST was encouraged and RT in the form of randomization testing was appropriate. During the analysis of the post-interviews the use of DST was documented and three distinct patterns of responses were discovered.

First, Jaylee, the teacher with the lowest pre-assessment score (PreA—1.00 out of 4.00, comfort-level [CL]—1.42 out of 5) and extremely low confidence with respect to statistical reasoning at the pre-interview, appeared to utilize DST with extreme proficiency. After visually comparing the distributions from the task with apparent confidence and skill, she turned immediately to the *Fathom2* environment to build the randomization test mechanism to conduct a randomization test to determine whether differences in means were statistically significant. She had already made a correct prediction. She navigated in an almost expert-like manner as she described all of the moves she was making. She exuded confidence in her computer skills and her statistical understanding that corresponded with her greatly improved post-assessment score (PostA—3.15, CL—3.43). She expressed that her background in computer science may have been helpful to her learning of *Fathom2*. Jaylee is an example of a teacher for whom DST appeared to be an important factor in supporting the improvement in understanding of *comparing distributions*. It was as if statistical reasoning and DST usage were synergistic and mutually complementary for Jaylee.

Second, June, a teacher with slightly better pre-assessment score (PreA—1.40, CL—1.57) than Jaylee and a similar low level of confidence with respect to statistical reasoning at the pre-interview, exhibited initial lack of confidence at the post-interview. When visually comparing the distributions, she did not appear confident in her responses, although she did correctly identify shapes and measures of center and variability. She incorrectly predicted whether the differences between means were statistically significant. When she moved to the DST environment, she elected to work with *Fathom2*. More detail is provided in Chapter V; however, once she began to navigate in *Fathom2*, it

was like the software environment was stimulating her to either remember or make connections between representations as she went along. Her expressed confidence appeared to grow as the interview progressed. She confessed to not working with the software during the professional development program because her partner dominated the computer, but was able to very adeptly navigate her way around. By the end of the interview, June's greatly improved post-assessment and comfort level scores appeared consistent with her interview responses (PostA—3.45, CL—3.57). June is an example of a teacher for whom the presence of DST in an RT environment may amplify statistical reasoning. There was a synergy between reasoning and technology usage, but it appeared that the use of technology was a catalyst driving the statistical reasoning.

Third, David is an A.P. Statistics teacher with relatively strong incoming understanding of and comfort-level with *comparing distributions* (PreA—3.15, CL—4.57). During the pre-interview and the professional development program, David exhibited a particularly procedural understanding of *comparing distributions*. He appeared to utilize *Fathom2* during the professional development program in appropriate ways and he contributed regularly to ongoing conversations, indicating he was finding value in the investigations involving DST and RT. During the post-interview, David was the only teacher who declined to engage with the technology to conduct a randomization test to compare the two distributions from the task. He expressed that he would need a lot of help to utilize either DST tool. Given his use of the tool during professional development, this response was surprising. Though he confidently and correctly compared the two distributions from the task using graphical representations and estimates of center and spread, he was not confident utilizing either DST. He claimed to

be confident that when he worked with the software again, his facility with it would return; he also claimed to be excited about using *Fathom2* with his students. David still exhibited a strong disposition toward procedural understanding of statistics, but his language had evolved to convey a greater appreciation of the value of multiple graphical representations and measures. His post-assessment and comfort-level scores increased (PostA—3.60, CL—4.71), he appeared proficient with the use of DST during professional development, he appeared confident and exhibited appropriate statistical reasoning without technology, but did not engage with DST during the interview. He may not have felt as competent with the technology as he did with the statistical concepts and he may have chosen to avoid showcasing what he may have felt was a weakness. Alternately, David may be a case where strong content knowledge inhibited the use of technology because the need to engage with the technology to make sense of the statistical ideas was not as necessary.

The three cases collectively represent different ways in which teachers appeared to utilize DST (*Fathom2* in these cases) and RT (randomization test in these cases) to compare distributions on a performance task in this study. In all three cases, teachers exhibited improvement in understanding of *comparing distributions* following the professional development program. Zbiek and colleagues (2007) suggest that “the core of instrumental genesis in mathematics education is understanding the mathematics of the technology and being able to use it for one’s own purposes” (p. 1179). From the perspective of instrumental genesis, it appears that *Fathom2* had become more of an instrument for Jaylee and June than it had for David. Even for Jaylee and June the way in which the artifact may have become instrumentalized appears notably different. These

differences, and potentially others, may be useful for professional development providers to consider when designing professional development programs with DST and RT with high school teachers.

This research suggests that the resampling techniques in a dynamic statistical tool environment have promise for supporting teachers' understanding of *comparing distributions*. Like the high school students using the permutation test (Barbella et al., 1990), high school teachers have been able to grasp the use of the randomization test. Similar to the results of Konold (1994), learners seemed to appreciate the opportunity to learn statistics with resampling approaches. In his study, after experiences with resampling approaches, learners exhibited a surprisingly weak understanding of the probabilistic ideas underpinning their work with sampling distributions. In this study, although there were instances of teachers with potential misconceptions about sampling distributions, the majority of teachers appeared to understand the hierarchy of objects (population, sample, collection of measures) that are a basis of reasoning from sampling distributions. The intentional way in which hierarchically structured objects and processes were coordinated through the use of DST and RT may help to explain why teachers' understanding of *comparing distributions*, especially sampling distributions, appeared to improve. Based on the recommendation of Saldanha (2004), this study offers possible insight into supporting learners' understanding of sampling distributions.

Question 3: Characteristics of Professional Development

The third research question was, "What characteristics of professional development for high school mathematics teachers contribute to their understanding of

comparing distributions?” The answer to this question is largely attributed to the analysis of the video of the professional development program as described in Chapter IV. The short answer to the question is that for professional development to contribute to teachers’ understanding of *comparing distributions*, it must be engineered so that teachers have the opportunity to explore *comparing distributions*. As conceived in this study, *comparing distributions* involved comparisons involving distributions using multiple representations and measures (e.g., Activity 1.3—Matching Plots to Variables, Activity 1.4—Matching Plots to Statistics), comparisons within and between distributions with respect to multiple measures of variability (e.g., Activity 1.4—Standard Deviation and Its Interpretation, Activity 3.5—Samples and Variation), and comparisons between sampling distributions under varying assumptions (e.g., Activities 1.2, 1.5, 2.3—Orbital Express, Activity 2.4—Random Rectangles, Activity 3.4—Is there a Relationship between Sample Size and the Sampling Distribution’s Standard Deviation?). Many of these comparisons were made flexibly with the use of dynamic statistics technology (DST). All of these comparisons were made in an environment where reasoning and justification were privileged over right answers, where engaging and rich tasks generated interest and enthusiasm for learning, where developing a shared understanding of concepts drove the direction of the session, and where technology was used as a tool to conjecture, test, and refine developing ideas and theories.

As the analysis of the emergent learning trajectory documented in Chapter IV illustrated, teachers’ participation in the professional development program evolved over the course of four short days. Six characteristics of professional development that emerged as influential in supporting teachers’ understanding of comparing distributions

were: (1) accessible and engaging investigations, (2) an intellectually safe environment, (3) sharing authority for knowledge, (4) making statistical arguments, (5) sense-making, and (6) the use of multiple representations. The last three characteristics were discussed in the previous section as they were also associated with the use of DST and RT.

Accessible and engaging investigations were seen as those with interesting contexts for teachers, those building on teachers' prior knowledge, those with connections to previous and future professional development activities, those with potential for use with high school students, and those that could be connected to teachers' current practice. An intellectually safe environment appeared to develop on the basis of all parties assuming a non-evaluative stance, accepting responsibility for sharing ideas and thinking with the group members and the class, agreeing that support for ideas and conjectures was balanced with the need for evidence, agreeing to listen carefully to others and share the airtime equitably among the group. The use of humor by all parties and willingness to express vulnerability, particularly with respect to mathematical and statistical content knowledge, were two additional contributors to establishing and maintaining the intellectually safe environment. Sharing authority for knowledge in the professional development classroom appeared to develop based upon the intentional and sustained press beyond "right answers," by agreeing to make thinking visible and public, by learning together and supporting and challenging each other, and by appreciating different points of view and experience.

From individuals with seemingly minimal statistical knowledge and a decidedly "received knowing" posture, by the end of four days, a collective intelligence within the community of learners had been discovered and nurtured at each site whereby ideas were

shared and built upon. Critiques were presented respectfully and challenges were not personal. Teachers' supported one another in their attempts to grapple with new technologies, new statistical concepts, and new norms of operation in a classroom. It was not perfect and there were missteps, but when viewing changes in knowing and understanding from the view of participation in social practices in which sense-making and reasoning are valued (Boaler & Greeno, 2000), a clear change in understanding of *comparing distributions* was seen. It would be overzealous to report that teachers made an abrupt transformation from received knowing to connected knowing; however, it is not likely too much of a stretch to claim that some, if not most teachers in this study, may have inched a little closer in that direction.

This study attempted to strengthen teachers' understanding of *comparing distributions* by putting teachers' thinking at the center of professional development (Mewborn, 2003). The professional development was designed and implemented utilizing recognized characteristics of high quality professional development (Loucks-Horsley et al., 2003; Weiss, 2006). It provided "an environment in which teachers could explore important ideas about data and statistics using new software tools" (Rubin et al., 2006, p. 19) and supported productive disciplinary engagement of high school mathematics teachers in a statistical learning environment (Engle & Conant, 2002). An attempt to coordinate the complexities associated with the many factors that may have helped to shape or contribute to teachers' understanding of comparing distributions is presented next.

Figure 39 depicts a model that emerged from the collective retrospective analyses that summarizes both the general content trajectory (the growing dots in the center of the

large arrow) and the main characteristics of the professional development that likely led to teachers' improved understanding of *comparing distributions* (around the perimeter of the central arrow). Whereas the statistical content evolved from the beginning of the lower left-hand end of the arrow toward the top right-hand end of the arrow, the characteristics of professional development were meant to be present throughout. The model is offered as a succinct representation which embodies the characteristics of professional development which seemed to contribute to teachers' evolving understanding of *comparing distributions*. For additional detail, see Table 17 in Chapter IV.

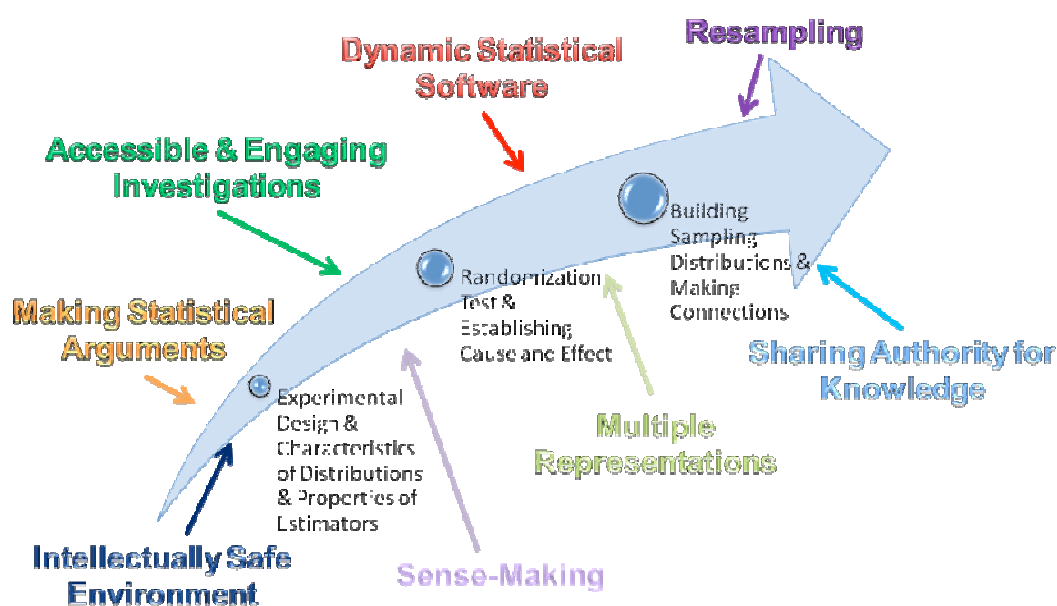


Figure 39. A model of teachers' evolving understanding of *comparing distributions*.

Discussion

The research reported here contributes to the literature in several ways. First, as Shaughnessy (2007) and others have discussed, there is tremendous need to support secondary teachers' understanding of statistics and much of the evidence of the current

state of knowledge is largely anecdotal. Shaughnessy specifically addresses the need for more research on students' conceptual growth in statistics when they work in technology rich environments, teachers' conceptions of statistics, classroom discourse in statistics, and teachers' beliefs and attitudes towards statistics. The present study chips away at pieces of each of those areas. It provides evidence to support the statement that high school mathematics teachers' understanding of some statistical ideas may not be strong, but also that under certain conditions, such as those described in this study, that understanding may be improved significantly. The research provides insight into the evolving statistical understanding of 56 high school mathematics teachers and suggests promising trajectories for future work, especially related to supporting understanding sampling distributions. As such, the research extends the sampling distribution work of other researchers (Chance et al., 2004; Lunsford et al., 2006; Saldanha, 2004; van der Meij & de Jong, 2006) by providing specific instructional suggestions for improving learners' understanding.

The use of resampling and dynamic statistical tools in a simulation environment, as described in this study, extends previous work with simulations and Monte Carlo techniques for supporting probability and statistical understanding of students through examining teachers experiences with them (Atkinson, 1975; Chance et al., 2000; Hecht, 1980; Lane & Peres, 2006; Shevokas, 1974; J. L. Simon, 1997; Travers, 1981; van der Meij & de Jong, 2006). Furthermore the study contributes to what we know about the effects of "landscape-type" data analysis tools (e.g., *Fathom2*) on learners' (in this case teachers') growth in statistical understanding and reasoning.

The research introduces several models (e.g., see Figures 38 and 39) with the potential to influence future design work in statistics or mathematics education research. A design principle that may be useful to others thinking about supporting teachers' work with dynamic technology software and simulations is discussed next.

Dynamic Technology Scaffolding: A Design Principle for Supporting Teachers' Understanding of Comparing Distributions

A particular finding from this research that emerged during the retrospective analysis and was not directly connected to any of the three research questions is the discovery of a design principle with potential for supporting teachers' understanding of important statistical ideas. I have chosen to call it *dynamic technology scaffolding*. In this case, *comparing distributions* was the important statistical idea, but the principle may be applicable more generally. In the paragraphs that follow, the principle will be described and explained in some detail.

Background for Dynamic Technology Scaffolding

As described previously, the randomization test was used as both a statistical tool and pedagogical device with the intent of supporting teachers' understanding of *comparing distributions*. Table 74 provides comparisons between the mechanics of hypothesis testing and randomization testing across the different technological environments utilized in this study.

Table 74

The Randomization Test Across Multiple Technological Environments

Hypothesis Testing Mechanics	Randomization Testing		
	Physical Simulation	<i>CPMP-Tools</i> (see Figure 40)	<i>Fathom2</i> (see Figure 41)
Collect experimental data	Write data values on cards	Enter data into <i>CPMP-Tools</i> (acts like a spreadsheet)	Enter data in <i>Fathom2</i> (acts like a spreadsheet)
Determine statistic of interest (calculate the difference between the measures of interest and record)	Determine statistic of interest (calculate the difference between the measures of interest and record)	Determine statistic of interest (select mean, median, or standard deviation)	Determine statistic of interest (use the formula feature to generate any measure of interest or difference between two measures)
Identify appropriate sampling distribution under the null hypothesis assumption	Put all cards together, representing the null hypothesis that there is no difference between the two populations of interest	Start animation Observe columns of data change	Stack the data (option to graph) New collection is generated TIER 1
	Shuffle the cards		Scramble the data (option to graph) New collection is generated May manually re-scramble the data TIER 2
	Distribute shuffled cards into piles of the same cardinality as the original data sets	Observe columns update (red and blue font indicates original group)	
	Compute the difference between the measures of interest	Observe difference between measure of interest display	Create a measure of the difference in statistics between two groups
	Record/accumulate result on a histogram—generates the empirical sampling distribution	Observe randomization distribution update	Collect measures New collection of measures is created Create histogram of measures TIER 3
	Repeat multiple times	Observe as animation continues the process; stop after desired number of trials	Use inspector to determine number of trials, select and collect more measures
Compare critical value	Examine where on the resultant distribution the original statistic is located	Locate the line on the randomization distribution representing the difference of interest from the original data	Plot original difference on the resultant randomization distribution graph
Determine p -value	Determine p -value based on the proportion of results equal to or more extreme than the original result	Determine p -value by summing the frequencies of the histogram bars at or beyond the original difference (may use the drag feature of the histogram to adjust bin-widths for ease of calculation)	Determine p -value by summing the frequencies of the histogram bars at or beyond the original difference (may use the drag feature of the histogram to adjust bin-width for ease of calculation)
Interpret the result	Determine if p -value is small enough to suggest that chance is not the likely explanation for the observed phenomenon (accept or reject the null hypothesis)		

Physical Simulation Environment

When comparing two samples, physical simulation is straight forward for relatively small data sets ($n_1, n_2 < 10$) using experimental data and index cards representing data values (Table 74). This simulation procedure was utilized during the Orbital Express activity, very early in the professional development program. The process encouraged an awareness of sampling variability throughout the resampling procedure. The physical procedure was important to ground learners' thinking about the kinds of samples produced under the assumption that there is no difference between populations, while at the same time allowing for comparison of shape, center, and spread of the resampled data. When all students work with the same sets of data, the results of the resampling trials are accumulated for the benefit of the whole class.

For all its pedagogical merits, physical simulation lacks efficiency. Shuffling, distributing, recalculating, and recording involve a good deal of time and effort for learners who may grow weary and become disinterested in this environment quickly. Even with a whole class effort, generating several hundred resamples can be time-consuming. After sufficient experience and conversation about the process and its interpretation, it may be advantageous for learners to move to an environment in which the process becomes automated but remains transparent. Hypothetically, such an environment might minimize time required to construct the distribution while preserving the essence of the process. The structure of the process may remain visible to the learner. The two different technological environments with this potential and utilized in the study were *CPMP-Tools* and *Fathom2*.

CPMP-Tools Environment

Along with other cognitive tools for mathematical and statistical investigation, *CPMP-Tools* contain a built-in feature specifically designed to conduct randomization tests. Building or growing the sampling distribution is as easy as clicking a button. Replication is a snap. The animation feature allows the learner to focus on the things that are changing and the aspects of the simulation that are remaining approximately invariant. Figure 40 provides a visual, non-dynamic view of the Randomization Distribution feature of *CPMP-Tools*. Data on the right side of the screen are represented in the first column as pooled data (the two original samples are noted in different colors on the screen); the second column contains a resample for one group (sorted by group); the third column contains a resample for the second group. The randomization distribution on the left represents the accumulation of 1000 resampled statistics (difference between the mean of the two groups in this case) with the vertical line on the graph representing the original statistic of interest.

With the animation feature enabled during the simulation, each time resampling is done, the right two columns of resampled data dynamically update, the statistic of interest (mean in this case) for each resample is calculated and shown at the bottom of each column, and the difference between the statistics is plotted on the histogram. In this example, the proportion of resampled statistics greater than or equal to the original test statistic is $4/1000 = 0.004$, indicating a very small likelihood that the original difference was due to chance. Hart et al. (2007) argue that the use of the randomization test in this technological environment may serve to amplify student learning of probability and statistics and support the conceptual understanding of statistical inference by students.

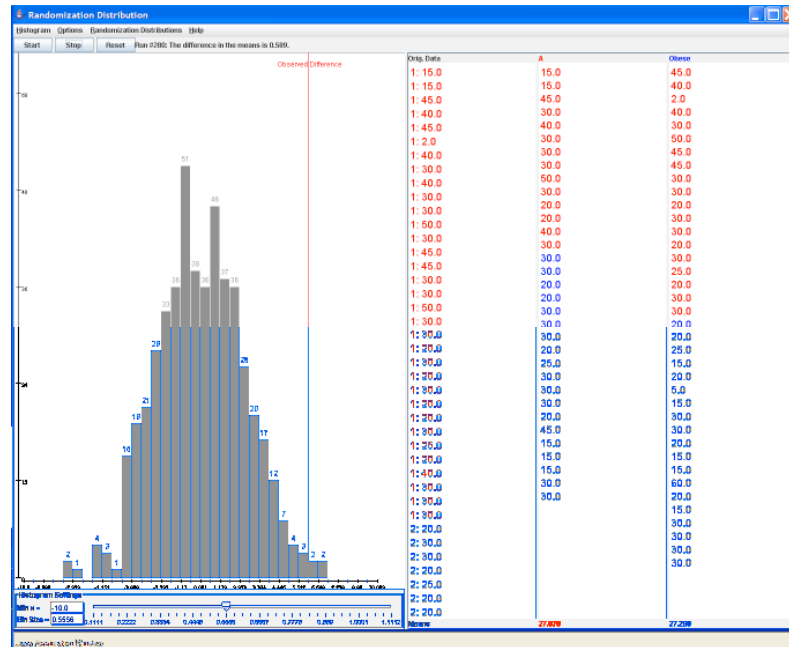


Figure 40. An example of the randomization distribution tool in *CPMP-Tools*.

Following one physical simulation and one brief experience with *CPMP-Tools* during the professional development program while conducting an investigation, teachers rated the ease of use of *CPMP-Tools* for conducting the randomization test on a scale from 1 (low) to 10 (high) with a mean 8.3 and median 9 ($n = 53$, min = 3, max = 10). At the same time, when asked to use the same scale to rate their current understanding of the randomization test, results were mean 6.6 and median 7 ($n = 55$, min = 1, max = 10). Together these data suggest the randomization test process was accessible to teachers and that *CPMP-Tools* was seen as easy to use. The dialogue seen in Chapter IV provides further illustration of the ways in which teachers began to interpret the randomization test and how generally receptive they were to this process.

Fathom2 Environment

In addition to developing improved statistical understanding, another goal of the professional development was to build teachers' capacity to utilize *Fathom2* software in order that teachers might support statistical inquiry within their classrooms. It was conjectured that once teachers understood the randomization test, a way to support their learning of *Fathom2* would be through their construction of the mechanisms in *Fathom2* to conduct the randomization test. One of the things that appeared quite important in this learning sequence was the ability to utilize the scaffolding afforded from the physical simulation and *CPMP-Tools* environments to invoke the actions from one or more environments and apply them in a new environment. A comparison of the learner demands when using *CPMP-Tools* versus *Fathom2* to conduct a randomization test is seen in Table 74. Table 74 highlights the active construction of mechanisms required to carry out the process in *Fathom2*. By contrast, in *CPMP-Tools*' Randomization Test environment, it may have been possible for the learner to passively observe and perhaps not fully understand the process, some evidence from written reflections and teacher interviews confirmed this speculation. It is conceivable that a learner may be able to use *CPMP-Tools* to conduct a randomization test and to correctly interpret the results without fully understanding or coordinating the process or the hierarchical objects. When asked on written reflection prompt 2A, "In your own words, describe what you have learned about the randomization test," some teachers indicated a sense of confusion. For example, one teacher wrote:

I learned how easy the computer makes it, but I am still foggy on the concepts. Would need much more practice & confidence to teach to kids. Fast w/many trials. (Reflect 2A, T82)

An example from the post interview suggested that for Jaylee, the *Fathom2* environment may have supported her understanding of the randomization test beyond that of *CPMP-*

Tools:

The Fathom more replicates what they [students] would do with the cards by hand, and so since their building it, it has more impact than having the computer just do it. You know this [CPMP-Tools] is fine once they [students] understand the process and now if I had a randomization, with the . . . I wouldn't have to necessarily you know, have built it myself, I would know how to read it. (PostI, Jaylee)

It was conjectured that the combination of tools and representations proved powerful in supporting learners' understanding of the randomization process and ultimately improving their understanding of *comparing distributions*.

Lane-Getaz's (2006) Simulation Process Model (SPM), a three-tier graphic pre-organizer to support students' connections between simulation activities and the logic of inference contains three tiers: (1) population (parameters), (2) samples (statistics), and (3) distribution of sample statistics. These tiers are consistent with mechanisms which may be seen in *CPMP-Tools* or developed in *Fathom2*. The mechanics of the randomization test process may be summarized as Tier 1—Display Data, Tier 2—Scramble Attributes, Tier 3—Collect Measures, all visible in Figure 41 and reinforced in Table 74.

Because of the construction demands in *Fathom2*, it is likely that understanding the entire randomization test process was a necessary condition for learners to successfully construct the mechanism. Though the end result of the simulation in *Fathom2* is nearly identical to that in *CPMP-Tools*, a number of additional dynamically-connected representations can be hot-linked in the construction. With the animation

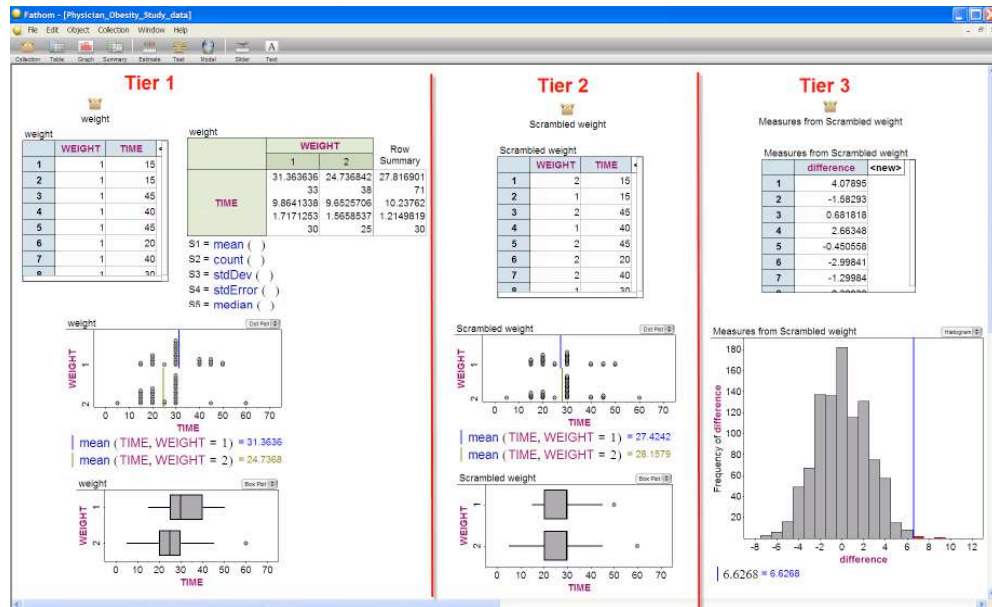


Figure 41. An example of the randomization distribution mechanism in *Fathom2*.

option turned on during simulation, objects in Tiers 2 and 3 dynamically update. Several things are afforded the learner in this environment:

1. Flexibility to investigate any statistic of interest (not limited to mean, median, and standard deviation as in *CPMP-Tools*);
2. Constructing a measure to generate a statistic of interest for comparison requires the coordination of fairly sophisticated mathematical symbolization (e.g., $difference = mean(time, weight = 1) - mean(time, weight = 2)$);
3. Scrambling attributes (Tier 2) and collecting measures (Tier 3) both generate new collections from which the hierarchical structure of sampling distributions may emerge;
4. Making connections between representations of histogram, boxplots, measures of center and spread, and the ways in which changes in one representation are reflected in the others;

5. Dynamically-linked representations make visible strong connection to issues of sampling variability under the null hypothesis of no difference between populations.

Evidence from a variety of sources suggested that the scaffolding of dynamic technology provided support for teachers' evolving understanding of *comparing distributions*. Teachers' written reflections strongly communicated the belief they were developing increasing facility with *Fathom2* as well as improving understanding of the randomization test. Teachers' self-reported understanding of the randomization test increased significantly from initial mean 6.46 and median 7 (min = 1, max = 10) on Day 2 to mean 8.28 and median 8.75 (min = 5, max = 10) by Day 4. The mean gain score of 1.63 was significantly greater than 0 ($t = 7.479, p < 0.0001$, effect size 1.01). Their reported comfort using *Fathom2* grew to from initial mean 6.68 and median 6.75 (min = 2, max = 10) on Day 2 to mean 7.30 and median 7 (min = 5, max = 10) by Day 4. The mean gain score of 0.625 was significantly greater than 0 ($t = 2.973, p = 0.002$, effect size 0.49). Teachers reported not feeling overly confident in their ability to use *Fathom2*; however, during a task-based portion of post-interviews, all nine teachers were able to describe the randomization test procedure and apply it to the task. Furthermore, six of the nine teachers elected to build the fairly complicated randomization test mechanism in *Fathom2* to successfully compare two distributions, even when *CPMP-Tools* was available. Two teachers elected to use *CPMP-Tools* and the remaining teacher confidently and correctly compared the distributions without the need to run a randomization test. As seen earlier in this chapter, teachers' content tests provided additional evidence that teachers' understanding of *comparing distributions* had

significantly improved. An example of what some teachers communicated about the value of *Fathom2* for conducting a randomization test is from Jordan:

I'd use Fathom over like . . . the Stat Plot tool [CPMP-Tools] just does it for you. So I guess for the . . . if I were going to do it with kids, I think they . . . I think I felt a better understanding when I went through the process of what I was doing, like when I had to set up the different steps. I felt like I understood what I was doing. When we first did it with Stat Plots [CPMP-Tools] I'm like, "Okay," but I didn't really grasp where we were getting things and what we were actually doing, but I think when I did it through the Fathom process, where it was pulling the samples and collecting the measures, then I felt like I had a better idea of what we were actually doing. (PostI, Jordan)

As a scaffold for learners wrestling with new statistical terrain as well as new demands from software environments, *CPMP-Tools* provided a vehicle which was elegant in its simplicity. It preserved the physical simulation actions while automating the process. As a cognitive tool, *CPMP-Tools* relieved the construction burden from the learner and allowed focus on the process and the result, potentially highlighting characteristics of the process without overwhelming the learner with technology demands or too many competing representations. According to Zbiek et al. (2007),

Cognitive tools play a special role in mathematical activity by externalizing representations (Heid, 1997). Through externalized, though limited, surrogates for a student's internal mental representations displayed on the surface of the screen, externalized representations become visible phenomena that can be shared and discussed with others (e.g., other learners or the teacher). By bringing such representations literally to the surface, a cognitive tool can allow for unique opportunities or exposing cognitive conflicts. (p. 1173)

Many statistics classes offer students opportunities to investigate probabilistic phenomena through the use of physical simulations and computer or graphing calculator simulations. At best, the research into the associated learning due to the use of simulations has been mixed (Mills, 2002). Some have argued that students are able to

play a passive role as they watch simulations (e.g., delMas et al., 1999; Lane & Peres, 2006). Others have seen some benefits through the use of integrated dynamically-linked simulations on student learning (van der Meij & de Jong, 2006). The research reported here suggests possible synergy through the intentional coordination of multiple tools.

Dynamic Technology Scaffolding: What Is It and Why Should We Care?

In addition to developing and applying the randomization test process in support of informal inference, a similar scaffolding of dynamic technology occurred through another sequence of learning activities for teachers. The sampling distribution of the mean was developed first through physical sampling from a finite population, and then demonstrated through a pre-constructed simulation using *Fathom2*, and finally with teachers' constructing the sampling distribution mechanism in *Fathom2*. The general model in Figure 42 illustrates the flow of activities and their collective relationship to the development of a central idea or concept.

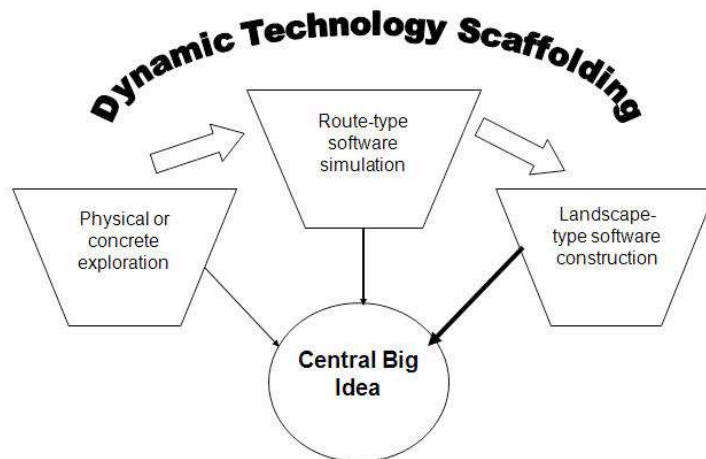


Figure 42. Model of dynamic technology scaffolding.

Beginning with a concrete or physical representation and investigation, the basis for a new concept is established. The dynamic technology may be rudimentary as is the case with physically shuffling and distributing index cards for the randomization test. A next level of activity with the potential to inform the development of the concept occurs via a structured simulation or demonstration in which the limitations imposed by the physical environment are removed and a cognitive tool is utilized to automate the process and assist the learner with noticing and connecting essential characteristics (e.g., *CPMP-Tools*, Java-applets, *Fathom2* simulations). At the third level of activity, the learner is actively engaged in the construction and manipulation of mathematical objects in a flexible technological environment (e.g., *Fathom2* or *Tinkerplots*) for the express purpose of engineering a mechanism which may possess similar attributes as the physical and/or simulated environment and one which will support additional investigation. Through the development of facility with a flexible, dynamic technological construction tool, the learner gains the additional benefit of the potential to further explore myriad relationships and conjectures. Each activity and associated dynamic technology serves as scaffolding for the development of further conceptual understanding at the next level. From one level to the next, the learner assumes more control over an increasingly sophisticated technological environment with the promise of supporting the conceptual development of important statistical or mathematical ideas and representations.

One of the catalysts for the emergence of this model occurred during professional development with the group of teachers at Site 1. The facilitator made an assumption about teachers' understanding and facility with technology that ultimately precipitated the proverbial "crying activity" (Activity 3.4). The investigation required teachers to

construct another sampling mechanism, seemingly “simpler” than the randomization mechanism, but due to the demands of the task and the relative novice *Fathom2* usage by teachers, the combination resulted in a near meltdown by at least one teacher and was supported by statements from many others. The zone of proximal development had been sufficiently overestimated by the facilitator. An adjustment to the learning trajectory for subsequent sessions confirmed that a demonstration simulation or the teacher use of a canned simulation to serve as a bridge between a concrete environment and abstract construction environment enabled teachers to successfully negotiate the demands of the construction task afterward.

This progression described is similar to the developmental modes of representation described by Bruner (1964): enactive (physical/concrete exploration), iconic (route-type software/simulation), and symbolic (landscape-type software construction). The model reflects the Vygotskian perspective that the use of tools may profoundly impact the ways in which learners come to understand statistical concepts and processes. *Dynamic technology scaffolding* is a design principle, emerging from this study, for developing and sequencing instructional tasks utilizing dynamic technology, with potential to support teachers’ conceptual understanding of challenging statistical (and likely non-statistical) ideas.

Discussion Summary

This research contributes to the literature related to statistics education for high school mathematics teachers. It contributes to the literature about potentially effective statistical content professional development for teachers with clear connections to

recommendations from professional education organizations in the United States. Furthermore, it contributes to advancing the knowledge of how learners' might utilize various forms of dynamic statistical tools and resampling techniques in support of improved statistical understanding. Finally, it extends and generates models for representing the evolution of content; proposed relationships among components of content, pedagogy, and technology; and suggests a design principle which may help to illuminate the myriad complexities of working in statistics education with practicing teachers.

Limitations and Implications

As with any research study, there are limitations. In the present study, one limitation refers to the target audience, who are high school mathematics teachers in relatively rural schools. It is possible that the research reported here may not generalize to other populations.

Another limitation in the study is that the researcher is also the designer as well as the facilitator of the professional development. Some may be skeptical of this decision and this was anticipated in advance of conducting the study. It was determined by the researcher and her dissertation committee that the opportunity to work with and study a large group of high school mathematics teachers while attempting to support their statistical understanding, as described herein, outweighed the cost of not doing the research. To aid in the transparency of the study, extensive data were collected, analyzed, and interpreted and a large sample of the various forms of data were presented for critical review. Unquestionably this type of research is best done with a team of people,

preferably a team of at least several researchers, but that option was not available for the dissertation work. Instead, one additional mathematics educator was present to videotape and conduct daily debriefing sessions of the professional development program with the primary researcher. He also provided reliability measures for the analysis of the content assessments and reacted to pieces of the written dissertation.

A third limitation concerns the facilitation skills of the researcher during the professional development. Some may argue that the improvement in teacher understanding documented in this research may simply be an artifact of a skilled and knowledgeable mathematics and statistics educator. Though arguably that line of reasoning may be valid, the hope would be that when conducting any type of professional development with teachers, the facilitator would be a skilled and knowledgeable other. That the professional development was “replicated” in three separate locations with distinctly different groups of teachers with very similar results, suggests that the design aspect of the intervention may have merit for continued research. Consistent, mostly positive signals were seen from each group of teachers, regardless of their location. That said, there is no guarantee that the improved teacher understanding of *comparing distributions* seen in the three professional development settings described here would occur in other settings. That similar findings occurred across three groups is promising, but preliminary at best.

A fifth limitation concerns the fact that this professional development was not specifically linked to teachers’ curricular work or upcoming work with students. Though the statistical content explored during the professional development was relevant for high school students and many of the instructional materials used during the session were

adapted from high school student materials, it was not the intent of this study to explore the impact of the teacher change on students. However, teachers received site licenses for *Fathom2* to use with their students as part of the grant supporting the larger professional development effort, thus the likelihood of their utilizing some of this work directly with students may be increased.

Finally, coming to understand what others' understand is a difficult business and any theories that have been espoused in this work are, in the words of Paul Cobb and colleagues (Cobb, Confrey, et al., 2003), "humble theories" limited by the constraints of the environment which situated the work.

On the positive side, this research improves our understanding of ways to deepen teachers' statistical content knowledge for teaching. It contributes to the knowledge base relative to design experiments in professional development settings. This research provides insight into how understanding of *comparing distributions* may evolve in a resampling, dynamic statistical tool environment and demonstrates that supporting teachers' statistical content knowledge is possible in a modest four-day technology-oriented, task-based professional development environment. Furthermore, it establishes a productive line of research regarding the tensions between use of technology, pedagogy, and statistics content utilizing the potential of resampling methods as vehicles to support the understanding of big ideas of statistics for both teachers and students.

Suggestions for Future Research

The design research reported here should be viewed as a first iteration of attempting to engineer a learning environment to support teachers' developing

understanding of *comparing distributions*. As seen previously, findings from this initial iteration have ramifications for theory and practice. It will be useful to continue to adjust and refine the design for use with other groups of teachers in different situations, particularly high school mathematics teachers from non-rural schools. It will be particularly important to attend to design characteristics to help ameliorate some of the potential misconceptions that some teachers exhibited at the end of the study. Because this study was with high school mathematics teachers, it may be useful to extend this work to mathematics pre-service teachers and study the relative affordances from a statistical experience as described in this study compared to an alternate required statistics experience for pre-service teachers.

Due to time constraints during this study, the use of the randomization test was limited to comparing two independent samples. Though it had been planned, through additional investigations, for teachers to explore matched pairs designs and bivariate correlations using randomization testing, time did not permit that experience. An extension of the work here might include those types of investigations in order to determine whether and what types of affordances might be made possible in support of teachers' understanding.

The design research described here suggests that under certain conditions, teachers' understanding of *comparing distributions* can evolve to fairly sophisticated levels in a relatively short period of time. Of particular interest might be how can we help teachers further connect the ideas from an environment as described in the present study to more of the theoretical underpinnings of statistics?

Because the nature of the statistical content addressed in this study is appropriate for high school students (College Board, 2006b; Franklin et al., 2007; NCTM, 2000), a reasonable extension of this work would be to investigate the ways in which experiences similar to those lived by teachers in this study may be lived by high school students. Dynamic technology scaffolding seemed important and relevant for teachers; would it be so for students?

Concluding Thoughts

It was an honor and a privilege to work closely with the teachers throughout the professional development work described in this study. Through their generosity of time and spirit, this work was able to unfold. I am forever humbled by teachers' willingness to share their thoughts and their lives as we collectively pursued the improvement of knowledge and understanding of statistical big ideas and the mechanisms that may support the improvement of that knowledge and understanding. Their voices and their experiences dominate this work. I personally experienced tremendous challenges and growth as a researcher and a practitioner throughout this experience. It is my greatest hope that our collective journey may continue to inform the work of others.

REFERENCES

- Abrahamson, D., & Wilensky, U. (2007). Learning axes and bridging tools in a technology-based design for statistics. *International Journal of Computers for Mathematical Learning*, 12(1), 23-55.
- Achieve, Inc. (2007). *Secondary mathematics benchmarks progressions: Grades 7-12*. Retrieved from <http://www.achieve.org/coursesequences>
- American Diploma Project. (2004). *Ready or not: Creating a high school diploma that counts*. Retrieved May 7, 2006, from <http://www.achieve.org>
- Atkinson, D. T. (1975). *A comparison of the teaching of statistical inference by Monte Carlo and analytic methods*. Urbana-Champaign: University of Illinois.
- Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2002). *The resilience of overgeneralization of knowledge about data representations*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Bakker, A. (2002). *Route-type and landscape-type software for learning statistical data analysis*. Paper presented at the International Conference on Teaching Statistics, Cape Town, South Africa.
- Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal*, 3(2), 64-83.
- Bakker, A., Biehler, R., & Konold, C. (2005). *Should young students learn about box plots?* Voorburg, The Netherlands: International Statistical Institute.
- Bakker, A., Derry, J., & Konold, C. (2006). *Using technology to support diagrammatic reasoning about center and variation*. Paper presented at the International Conference on Teaching Statistics (ICOTS-7), Salvador, Bahia, Brazil.
- Bakker, A., & Frederickson, A. (2005). Comparing distributions and growing samples by hand and with a computer tool. In W. J. Masalski & P. C. Elliot (Eds.), *Technology-supported mathematics learning environments* (NCTM 67th Yearbook) (pp. 75-92). Reston, VA: National Council of Teachers of Mathematics.
- Bakker, A., & Gravemeijer, K. P. E. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking*. Boston: Kluwer Academic.

- Ball, D. L. (2000). Bridging practices: Intertwining content and pedagogy in teaching and learning to teach. *Journal of Teacher Education*, 51, 241-247.
- Ball, D. L. (2002). Knowing mathematics for teaching: Relations between research and practice. *Mathematics and Education Reform Newsletter*, 14, 1-5.
- Ball, D. L. (2003). *What mathematical knowledge is needed for teaching mathematics?* Retrieved May 30, 2006, from <http://www-personal.umich.edu/~dball/Presentations/BallMathSummitFeb03.pdf>
- Ball, D. L., & Bass, H. (2000). Interweaving content and pedagogy in teaching and learning to teach: Knowing and using mathematics. In J. Boaler (Ed.), *Multiple perspectives on mathematics teaching and learning: International perspectives on mathematics education* (pp. 83-104). Westport, CT: Ablex.
- Barbella, P., Denby, L., & Landwehr, J. M. (1990, February). Beyond exploratory data analysis: The randomization test. *Mathematics Teacher*, 83, 144-149.
- Bartoszynski, R., & Niewiadomska-Bugaj, M. (1996). *Probability and statistical inference*. New York: John Wiley & Sons.
- Belenky, M. F., Clinchy, B. M., Goldberger, N. R., & Tarule, J. M. (1986). *Women's ways of knowing: The development of self, voice and mind*. New York: Basic Books.
- Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal*, 3, 42-63.
- Ben-Zvi, D., & Garfield, J. (Eds.). (2004a). *The challenge of developing statistical literacy, reasoning and thinking*. Norwell, MA: Kluwer Academic.
- Ben-Zvi, D., & Garfield, J. (2004b). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3-15). Boston: Kluwer Academic.
- Biehler, R. (2001). "Girls (tend to) watch less television than boys" — Students' hypotheses and data exploration strategies in group comparison tasks. Presentation at LOGOS #10, Mathematics Education Unit, Department of Mathematics, The University of Auckland, New Zealand.
- Biehler, R., & Kombrink, K. (1999). Der datendetektiv auf spurensuche. *Mathematische lehren*, 97, 23-46.
- Boaler, J., & Greeno, J. (2000). Identity, agency, and knowing in mathematical worlds. In J. Boaler (Ed.), *Multiple perspectives on mathematics teaching and learning: International perspectives on mathematics education* (pp. 171-200). Westport, CT: Ablex.

- Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher*, 33(8), 3-15.
- Brandom, R. B. (2000). *Articulating reasons: An introduction to inferentialism*. Cambridge, MA: Harvard University Press.
- Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.
- Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences*, 2(2), 141-178.
- Bruner, J. S. (1964). The course of cognitive growth. *American Psychologist*, 19, 1-15.
- Canada, D. (2006). Elementary pre-service teachers' conceptions of variation in a probability context. *Statistics Education Research Journal*, 5(1), 36-63.
- Canada, D., & Makar, K. (2006). *Preservice teachers' informal descriptions of variation*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295-323). Boston: Kluwer Academic.
- Chance, B., Garfield, J., & delMas, R. (2000). *Developing simulation activities to improve students' statistical reasoning*. Paper presented at the International Conference on Technology in Mathematics Education, Auckland, NZ.
- Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, 104(9), 801-823.
- Cobb, P. (1999). Individual and collective mathematical development: The case of statistical data analysis. *Mathematical Thinking and Learning*, 1(1), 5-43.
- Cobb, P. (2000). The importance of a situated view of learning to the design of research and instruction. In J. Boaler (Ed.), *Multiple perspectives on mathematics teaching and learning* (pp. 45-82). Westport, CT: Ablex.
- Cobb, P., Confrey, J., DiSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, 32(1), 9-13.
- Cobb, P., & McClain, K. (2004). Principles of instructional design for supporting the development of students' statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 375-395). Dordrecht, The Netherlands: Kluwer Academic.

- Cobb, P., McClain, K., & Gravemeijer, K. (2003). Learning about statistical covariation. *Cognition and Instruction*, 2(1), 1-78.
- Cobb, P., & Yackel, E. (1996). Constructivist, emergent, and sociocultural perspectives in the context of developmental research. *Educational Psychologist*, 31, 175-190.
- College Board. (2006a). *AP examination volume changes (1995-2005)*. Retrieved May 7, 2006, from http://apcentral.collegeboard.com/repository/examvolumechange_47028.pdf
- College Board. (2006b). *College Board standards for college success: Mathematics and statistics*. Retrieved from http://www.collegeboard.com/prod_downloads/about/association/academic/mathematics-statistics_cbscs.pdf
- Conference Board of the Mathematical Science (CBMS). (1982). *The mathematical sciences curriculum K-12: What is still fundamental and what is not*. Report to NSB Commission on Precollege Education in Mathematics, Science, and Technology. Washington, DC: Author.
- Conference Board of the Mathematical Science (CBMS). (2001). *The mathematical education of teachers*. Providence, RI: American Mathematical Society.
- Coxford, A. F., Fey, J. T., Hirsch, C. R., Schoen, H. L., Hart, E. W., Keller, B. A., et al. (2003). *Contemporary mathematics in context: A unified approach*. New York: Glencoe McGraw-Hill.
- 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).
- Edgington, E. S. (1995). *Randomization tests* (3rd ed.). New York: Marcel Dekker.
- Engle, R., & Conant, F. (2002). Guiding principles for fostering productive disciplinary engagement: Explaining emerging argument in a community of learners classroom. *Cognition and Instruction*, 20(4), 399-483.
- Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., et al. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-K-12 curriculum framework*. Alexandria, VA: American Statistical Association.
- Frykholm, J. (2004). Teachers' tolerance for discomfort: Implications for curricular reform in mathematics. *Journal of Curriculum and Supervision*, 19(2), 125-149.
- Gal, I., Rothschild, K., & Wagner, D. A. (1989). *Which group is better? The development of statistical reasoning in school children*. Paper presented at the meeting of the Society for Research in Child Development, Kansas City, MO.

- Gardner, D. P., Larsen, Y. W., Baker, W. O., Campbell, A., Crosby, E. A., Foster, C. A., Jr., et al. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: National Commission for Excellence in Education. (ERIC Document Reproduction Service No. ED 226 006)
- Garfield, J., & Ben-Zvi, D. (2004). Research on statistical literacy, reasoning, and thinking: Issues, challenges, and implications. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 397-409). Boston: Kluwer Academic.
- Garfield, J., & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. *Statistics Education Research Journal*, 4(1), 92-99.
- Goldin, G. A. (2003). Representation in school mathematics: A unifying research perspective. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to Principles and Standards for School Mathematics* (pp. 275-285). Reston, VA: National Council of Teachers of Mathematics.
- Good, P. I. (1999). *Resampling methods: A practical guide to data analysis*. Boston: Birkhauser.
- Gould, R. (2004). Variability: One statistician's view. *Statistics Education Research Journal*, 3(2), 7-16.
- Gravemeijer, K., & Bakker, A. (2006). *Design research and design heuristics in statistics education*. Paper presented at the International Conference on Teaching Statistics-ICOTS 7, Salvador, Bahia, Brazil.
- Greeno, J. G. (2003). Situative research relevant to standards for school mathematics. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to Principles and Standards for School Mathematics* (pp. 304-332). Reston, VA: National Council of Teachers of Mathematics.
- Hammerman, J., & Rubin, A. (2004). Strategies for managing statistical complexity with new software tools. *Statistics Education Research Journal*, 3(2), 17-41.
- Hammerman, J., & Rubin, A. (2006). *Saying the same (or a different) thing: How shape affects ideas about distribution in a software exploration environment*. Paper presented at the International Conference on Teaching Statistics-ICOTS 7, Salvador, Bahia, Brazil.
- Hart, E. W., Hirsch, C. R., & Keller, S. A. (2007). Amplifying student learning in mathematics using curriculum-embedded, Java-based software. In W. G. Martin & M. E. Strutchens (Eds.), *The learning of mathematics: Sixty-ninth yearbook* (pp. 175-202). Reston, VA: National Council of Teachers of Mathematics.

- Hecht, J. E. (1980). *Using the Monte Carlo method to teach probabilistic problem solving to ninth grade general mathematics students*. Urbana-Champaign: University of Illinois.
- Heid, M. K. (1997). The technological revolution and the reform of school mathematics. *American Journal of Education*, 106, 5-61.
- Hesterberg, T. (2006). Bootstrapping students' understanding of statistical concepts. In G. F. Burrill (Ed.), *Thinking and reasoning with data and chance: Sixty-eighth yearbook* (pp. 391-416). Reston, VA: National Council of Teachers of Mathematics.
- Hirsch, C. R., Fey, J. T., Hart, E. W., Schoen, H. L., & Watkins, A. E. (2008). *Core-Plus Mathematics: Contemporary mathematics in context* (2nd ed.) New York: Glencoe McGraw-Hill.
- Jones, P. S., & A. F. Coxford, J. (1970). The goals of history: Issues and forces. In *A history of mathematics education in the United States and Canada. 32nd yearbook* (pp. 1-7). Washington, DC: National Council of Teachers of Mathematics.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430-454.
- Keller, S. (2006). *CPMP-Tools* (Version 1.1, trademark pending). East Lansing, MI: Michigan State University and Core-Plus Mathematics Project.
- Key Curriculum Press. (2005a). *FathomTM Dynamic Statistics Software* (Version 2.0) [Computer software]. Emeryville, CA: Author.
- Key Curriculum Press. (2005b). *Tinkerplots Dynamic Data Exploration* (Version 1.0) [Computer software]. Emeryville, CA: Author.
- Konold, C. (1994). *Understanding probability and statistical inference through resampling*. Paper presented at the First Scientific Meeting of the IASE, Universita de Perugia, Italy.
- Konold, C., & Higgins, T. L. (2003). Reasoning about data. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to Principles and Standards for School Mathematics*. Reston, VA: National Council of Teachers of Mathematics.
- Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, 33(4), 259-289.
- Konold, C., Pollatsek, A., Well, A., & Gagnon, A. (1997). Students analyzing data: Research of critical barriers. In J. B. Garfield & G. Burrill (Eds.), *Research on the*

- role of technology in teaching and learning statistics: 1996 Proceedings of the 1996 IASE Round Table Conference* (pp. 151-167). Voorburg, The Netherlands: International Statistical Institute.
- Krauss, S., & Wassner, C. (2002). *How significance tests should be presented to avoid the typical misinterpretations*. Paper presented at the Sixth International Conference on Teaching Statistics, Cape Town, South Africa.
- Lane, D. M., & Peres, S. C. (2006). *Interactive simulations in the teaching of statistics: Promise and pitfalls*. Paper presented at the Seventh International Conference on Teaching Statistics, Salvador, Bahia, Brazil.
- Lane-Getaz, S. J. (2006). What is statistical thinking and how is it developed? In G. F. Burrill & P. C. Elliott (Eds.), *Thinking and reasoning with data and chance: Sixty-eighth yearbook* (Vol. 68, pp. 273-290). Reston, VA: National Council of Teachers of Mathematics.
- Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998). *Connected mathematics*. Menlo Park, CA: Dale Seymour.
- Lehrer, R., & Schauble, L. (2004). Modeling natural variation through distribution. *American Educational Research Journal*, 41(3), 635-680.
- Lerman, S. (2000). The social turn in mathematics education research. In J. Boaler (Ed.), *Multiple perspectives on mathematics teaching and learning* (pp. 19-44). Westport, CT: Ablex.
- Loucks-Horsley, S., Love, N., Stiles, K. E., Mundry, S., & Hewson, P. W. (2003). *Designing professional development for teachers of science and mathematics* (2nd ed.). Thousand Oaks, CA: Corwin Press.
- Loucks-Horsley, S., Stiles, K., & Hewson, P. (1996). Principles of effective professional development for mathematics and science education: A synthesis of standards. *NISE Brief*, 1(1).
- Lunsford, M. L., Rowell, G. H., & Goodson-Epsy, T. (2006). Classroom research: Assessment of student understanding of sampling distributions of means and the Central Limit Theorem in post-calculus probability and statistics classes. *Journal of Statistics Education*, 14(3).
- Makar, K. (2004). *Developing statistical inquiry: Prospective secondary mathematics and science teachers' investigations of equity and fairness through analysis of accountability data*. Austin: University of Texas.
- 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.

- Makar, K., & Confrey, J. (2004). Secondary teachers' statistical reasoning in comparing two groups. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 353-374). Boston: Kluwer Academic.
- Makar, K., & Confrey, J. (2005). "Variation-talk": Articulating meaning in statistics. *Statistics Education Research Journal*, 4(1), 27-54.
- McClain, K., Cobb, P., & Gravemeijer, K. (2000). Supporting students' ways of reasoning about data. In M. J. Burke & F. R. Curcio (Eds.), *Learning mathematics for a new century: 2000 yearbook* (pp. 174-188). Reston, VA: National Council of Teachers of Mathematics.
- Meletiou-Mavrotheris, M., & Lee, C. (2002). Teaching students the stochastic nature of statistical concepts in an introductory statistics course. *Statistics Education Research Journal*, 1(2), 22-37.
- Meletiou-Mavrotheris, M., & Lee, C. (2003). *Studying the evolution of students' conceptions of variation using the transformative and conjecture-driven research design* [CD-ROM]. Paper presented at the Third International Research Forum on Statistical Reasoning, Thinking, and Literacy, Lincoln, NE.
- Mewborn, D. S. (2003). Teaching, teachers' knowledge, and their professional development. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to Principles and Standards for School Mathematics* (pp. 45-52). Reston, VA: National Council of Teachers of Mathematics.
- Mickelson, W. T., & Heaton, R. M. (2003, July). *Purposeful statistical investigation merged with K-6 content: Variability, learning and teacher knowledge used in teaching* [CD-ROM]. Paper presented at the Third International Research Forum on Statistical Reasoning, Thinking, and Literacy Lincoln, NE.
- Mickelson, W. T., & Heaton, R. M. (2004). Primary teachers' statistical reasoning about data. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 327-352). Boston: Kluwer Academic.
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.
- Mills, J. D. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education*, 10(1).
- Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education*, 26, 20-39.
- Mooney, E. S. (2002). A framework for characterizing middle school students' statistical thinking. *Mathematical Thinking and Learning*, 4(1), 23-63.

- Moore, D. S. (1990). Uncertainty. In L. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95-137). Washington, DC: National Academy Press.
- Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review*, 65(2), 123-137.
- Moore, D. S., & McCabe, G. P. (2005). Bootstrap methods and permutation tests. In D. S. Moore & G. P. McCabe (Eds.), *Introduction to the practice of statistics* (5th ed.). New York: W. H. Freeman.
- National Commission on Mathematics and Science Teaching. (2000). *Before it's too late*. Washington, DC: Author.
- National Committee on Mathematical Requirements. (1923). *The reorganization of mathematics in secondary education*. New York: Mathematical Association of America.
- National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press.
- National Research Council. (2005). *How students learn: Mathematics in the classroom*. Washington, DC: National Academies Press.
- National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.
- National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.
- Noddings, N., vonGlaserfeld, E., Goldin, G., Baroody, A. J., Ginsburg, H. P., Davis, R. B., et al. (1990). *Constructivism in mathematics education*. Reston, VA: National Council of Teachers of Mathematics.
- Petrosino, A. J., Lehrer, R., & Schauble, L. (2003). Structuring error and experimental variation as distribution in the fourth grade. *Mathematical thinking and learning*, 5(2 & 3), 131-156.
- Pfannkuch, M. (2006). *Informal inferential reasoning*. Paper presented at the International Conference on Teaching Statistics-ICOTS 7, Salvador, Bahia, Brazil.
- Pfannkuch, M., & Brown, C. (1996). Building on and challenging students' intuitions about probability: Can we improve undergraduate learning? *Journal of Statistics Education*, 4(1).
- Pfannkuch, M., Budgett, S., Parsonage, R., & Horring, J. (2004, July). *Comparison of data plots: Building a pedagogical framework*. Paper presented at the Tenth International Congress on Mathematics Education (ICME-10), Copenhagen, Denmark.

- Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher*, 29(1), 4-15.
- Reading, C. (2004). Student description of variation while working with weather data. *Statistics Education Research Journal*, 3(2), 84-105.
- Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 201-226). Boston: Kluwer Academic.
- Rubin, A., Bruce, B., & Tenney, Y. (1990). *Learning about sampling: Trouble at the core of statistics*. Paper presented at the International Conference on Teaching Statistics (ICOTS3), Dunedin, New Zealand.
- Rubin, A., & Hammerman, J. K. (2006). Understanding data through new software representations. In G. Burrill (Ed.), *Thinking and reasoning with data and chance: Sixty-eighth yearbook* (Vol. 68, pp. 241-256). Reston, VA: National Council of Teachers of Mathematics.
- Rubin, A., Hammerman, J. K. L., & Konold, C. (2006). *Exploring informal inference with interactive visualization software*. Paper presented at the International Conference on Teaching Statistics-ICOTS 7, Salvador, Bahia, Brazil
- Russell, S. J., & Mokros, J. R. (1990). *What's typical? Children's and teachers' ideas about average*. Paper presented at the International Conference on Teaching Statistics (ICOTS3), Dunedin, New Zealand.
- Saldanha, L. (2004). *"Is this sample unusual?": An investigation of students exploring connections between sampling distributions and statistical inference*. Nashville, TN: Vanderbilt University.
- Saldanha, L., & Thompson, P. (2002, October). *Students' scheme-based conceptions of sampling and its relationship to statistical inference*. Paper presented at the annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Athens, GA.
- Schoenfeld, A. H. (2007). Method. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (Vol. 1, pp. 69-107). Charlotte, NC: Information Age.
- Shaughnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 465-494). New York: MacMillan.
- Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (Vol. 2, pp. 957-1010). Charlotte, NC: Information Age.

- Shaughnessy, J. M., Ciancetta, M., & Canada, D. (2004, July). *Types of student reasoning on sampling tasks*. Paper presented at the 28th International Group for the Psychology of Mathematics Education, Bergen, Norway.
- Shevokas, C. (1974). *Using a computer-oriented Monte Carlo approach to teach probability and statistics in a community college general mathematics course*. Urbana-Champaign: University of Illinois.
- Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2), 4-14.
- Simon, J. L. (1994). What some puzzling problems teach about the theory of simulation and the use of resampling. *The American Statistician*, 48(4), 290-293.
- Simon, J. L. (1997). *Resampling: The new statistics* (2nd ed.). Duxbury, MA: Thompson International.
- Simon, J. L., Atkinson, D. T., & Shevokas, C. (1976). Probability and statistics: Experimental results of a radically different teaching method. *The American Mathematical Monthly*, 83(9), 733-739.
- Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education*, 26(2), 114-145.
- Simon, M. A., & Tzur, R. (2004). Explicating the role of mathematical tasks in conceptual learning: An elaboration of the hypothetical learning trajectory. *Mathematical Thinking and Learning*, 6(2), 91-104.
- Stein, M. K., Smith, M. S., Henningsen, M., & Silver, E. A. (2000). *Implementing standards-based mathematics instruction: A casebook for professional development*. New York: Teachers College Press.
- Sterling, J., & Gray, M. W. (1991). The effect of simulation software on students' attitudes and understanding in introductory statistics. *Journal of Computers in Mathematics and Science Teaching*, 10(4), 51-56.
- Stigler, J. W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. New York: The Free Press.
- Strauss, A. L., & Corbin, J. M. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Thousand Oaks, CA: Sage.
- Torok, R., & Watson, J. (2000). Development of the concept of statistical variation: An exploratory study. *Mathematics Education Research Journal*, 12(2), 147-169.
- Travers, K. J. (Ed.). (1981). *Using Monte Carlo methods to teach probability and statistics*. Reston, VA: National Council of Teachers of Mathematics.

- Tukey, J. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.
- Usiskin, Z. (2003). A personal history of the UCSMP secondary school curriculum, 1960-1999. In G. M. A. Stanic & J. Kilpatrick (Eds.), *A history of school mathematics* (Vol. 1, pp. 673-736). Reston, VA: National Council of Teachers of Mathematics.
- Van De Walle, J. A. (2004). *Elementary and middle school mathematics: Teaching developmentally*. Boston: Pearson Education/Allyn & Bacon.
- van der Meij, J., & de Jong, T. (2006). Supporting students' learning with multiple representations in a dynamic simulation-based learning environment. *Learning and Instruction*, 16(3), 199-212.
- vonGlaserfeld, E. (1990). An exposition of constructivism: Why some like it radical. In R. B. Davis, C. A. Maher, & N. Noddings (Eds.), *Constructivist views on the teaching and learning of mathematics* (pp. 19-30). Reston, VA: National Council of Teachers of Mathematics.
- vonGlaserfeld, E. (1995). *Radical constructivism: A way of knowing and learning*. Washington, DC: Falmer Press.
- Wagner-Krankel, M. E. (1990). *Misconceptions of probability: An experiment with an activity-based, model-building approach using computer simulation Monte Carlo techniques to teach introductory probability at the college level*. Austin: University of Texas.
- Watson, J. M. (2004). Developing reasoning about samples. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 277-294). Boston: Kluwer Academic.
- Watson, J. M., Kelly, B. A., Callingham, R. A., & Shaughnessy, M. (2003). The measurement of school students' understanding of statistical variation. *International Journal of Mathematical Education in Science and Technology*, 34(1), 1-29.
- Watson, J. M., & Moritz, J. B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics*, 37, 145-168.
- Watson, J. M., & Moritz, J. B. (2000). Developing concepts of sampling. *Journal for Research in Mathematics Education*, 31(1), 44-70.
- Weiss, I. (2006). Bridging the gap between research and practice in professional development. Retrieved from www.ccsso.org/content/PDFs/ResearchonEffectivePDfinal.pdf

- Well, A., Pollatsek, A., & Boyce, S. J. (1990). Understanding the effects of sample size on the variability of the mean. *Organizational Behavior and Human Decision Processes*, 47, 289-312.
- Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, 67, 223-265.
- Wilkins, J. L. M. (1999). The cereal box problem revisited. *School Science and Mathematics*, 99(3), 117-123.
- Wood, M. (2005). The role of simulation approaches in statistics. *Journal of Statistics Education*, 13(3).
- Yackel, E., & Cobb, P. (1996). Sociomathematical norms, argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education*, 27(4), 458-477.
- Zbiek, R. M., Heid, M. K., Blume, G. W., & Dick, T. P. (2007). Research on technology in mathematics education. In Frank K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (Vol. 2, pp. 1169-1208). Charlotte, NC: Information Age.
- Zitek, E. (2006). *Physicians' reactions to patient size*. Retrieved May 20, 2006, from http://onlinestatbook.com/case_studies/weight.html

Appendix A
Pre-Assessment

☐

Check this box if you
do **NOT** agree to
allow use of your
responses for
research purposes

Code # _____

High School Mathematics Teachers' Evolving Understanding of Comparing Distributions

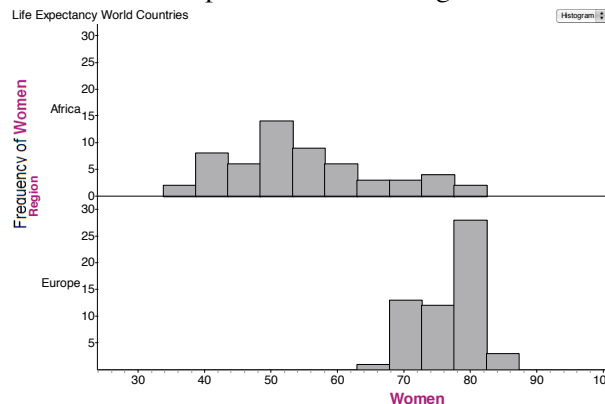
Summer 2006

Statistics Background. Please describe the statistics coursework/experiences you've had outside of M²RI. Include major topics studied. Also, have you taught a probability, statistics, or A. P. statistics course?

Statistics Comfort-level. Please rate your level of comfort with each topic listed below by circling the level that best corresponds to a rating with **1 being very low/none** and **5 being high comfort**:

	low comfort		med comfort		high comfort
Descriptive statistics (mean, standard deviation, z-score)	1	2	3	4	5
Statistical Graphs (histogram, boxplot, bar graph)	1	2	3	4	5
Distributions (normal, chi-square, probability density functions)	1	2	3	4	5
Experimental Design (surveys, blocking, bias, sampling methods)	1	2	3	4	5
Correlation and Regression (least squares, r^2 , residuals, outliers)	1	2	3	4	5
Sampling Distributions (Central Limit Theorem)	1	2	3	4	5
Statistical Inference (t-tests, confidence intervals, chi-square tests, power, Type II error, ANOVA)	1	2	3	4	5

1. The following distributions represent average life expectancies for women from Africa and Europe. These life expectancies are computed for various regions within each country.



(Source: *Fathom2*; Taken from www.overpopulation.com which cites McDevitt, Thomas M., World population profile: 1998. Washington, DC: U.S. Government Printing Office, 1999)

How are the distributions of life expectancies for African women and European women similar?

How are the distributions of life expectancies for African women and European women different?

What can you say about the life expectancies of women from Africa and Europe? Please be specific.

2. Given the average summer temperature in cities P and Q, explain briefly how you would decide which of the following two events is more unusual: a 90 degree summer day in city P or a 90 degree summer day in city Q.

3. A certain town has two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50%, sometimes lower. For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days?

- A) The larger hospital
- B) The smaller hospital
- C) About the same number of days (within 5% of each other)
- D) Can't tell

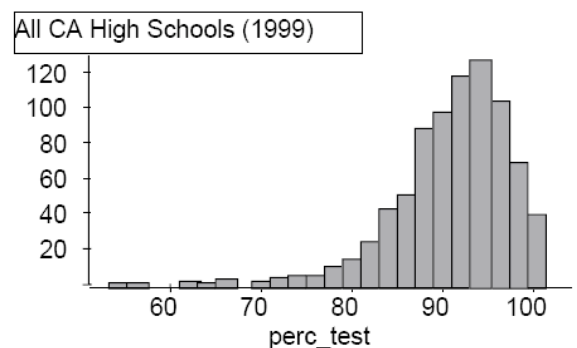
4. The **graph below right** shows the distribution of the percent of students tested on the state assessment at each of 806 high schools in California. Circle TRUE, FALSE, or CAN'T TELL **for each of the following statements.**

- a. The distribution is skewed right. TRUE FALSE CAN'T TELL
- b. The median is greater than the mean. TRUE FALSE CAN'T TELL

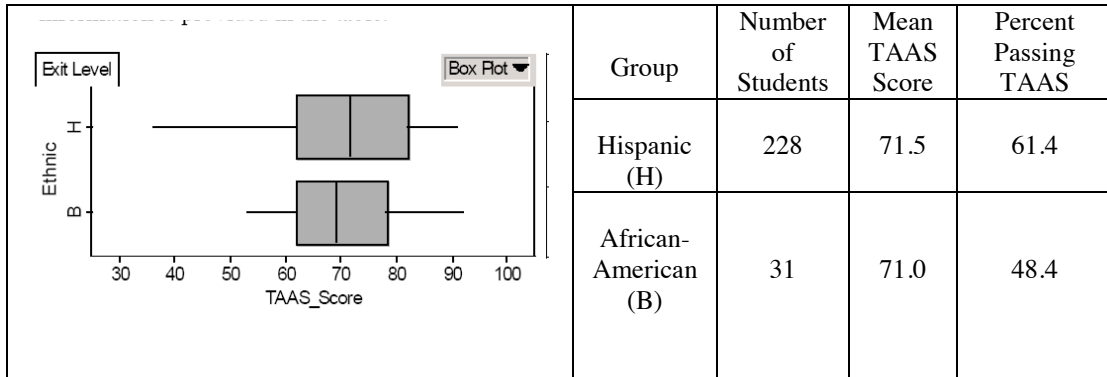
5. In the graph at right, how many schools are above the mean?

Choose one:

- A) Exactly half of the schools are above the mean.
- B) More than half of the schools are above the mean.
- C) Less than half of the schools are above the mean.
- D) I cannot answer the question without calculating the proportion of schools above the mean.



6. The pair of boxplots below represent the performance on the 2000 Texas state TAAS exam of two groups of 10th-grade students at an urban high school. The top boxplot describes the performance of 228 Hispanic students while the bottom boxplot represents the performance of the 31 African-American students. The school is considered “low-performing” if less than 50% of the students in any subgroup pass the exam. A score of 70 is considered passing. Additional information is provided in the table.



List three conclusions that would complete the following sentence: “By comparing the performance of Hispanic students with the performance of African-American students, I would draw the following conclusions...”

1)

2)

3)

7. When a set of data has suspect outliers, which of the following are preferred measures of center and of variability for describing the distribution?

- A) mean and standard deviation
- B) mean and variance
- C) mean and range
- D) median and range
- E) median and interquartile range
- F) can't tell

8.



(Image from kelloggs.com/promotions/cars/racers.shtml, retrieved June 9, 2006. Problem adapted from Wilkins, J. L. M. (1999). The cereal box problem revisited. *School Science and Mathematics*, 99(3), 117-123.)

Inside each box of specially marked boxes of Kellogg's® cereals, you will find one of six possible free *Cars* RACERS. Keegan really wants to collect all six RACERS and is trying to convince his parents to buy enough boxes of cereal in order to collect the set. His mother, with some knowledge of probability, suggests that he simulate the purchase of boxes of cereal in order to determine about how many boxes he may expect to purchase. Assuming equal chance of getting any of the six RACERS with one purchase, Mom suggests that Keegan use a die with each face of the die representing one RACER to simulate buying boxes of cereal. She asks Keegan to keep track of the total number of rolls of the die needed to “collect” one of each of the cars (numbers 1, 2, 3, 4, 5, and 6). Keegan conducted this simulation 5 times and the results of his trials are below.



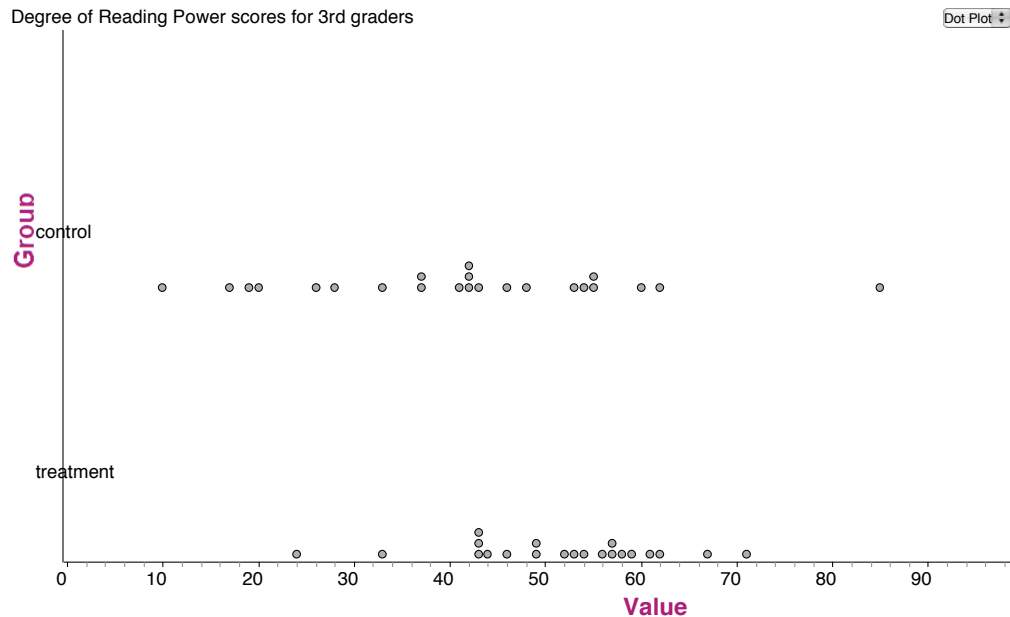
Results from Five Trials Using a Six-Sided Die to Represent the Different RACERS Vehicles

TRIALS	Prizes (sides of die)						N
	1	2	3	4	5	6	
1	///	///	/	///	////	////	19
2	///	//// //	/	///	///	///	21
3	//// /	/	/	/	/	//	12
4	//	/	///	//	/	/	10
5	/	////	/	//	//// /	//	16

What can you say, based upon Keegan's simulation, about the number of boxes of cereal he should expect to buy in order to collect the entire RACERS set? Please explain your reasoning.

9. A study was conducted to determine whether new “directed reading activities” improve the reading ability of elementary school students, as measured by their Degree of Reading Power (DRP) scores. The study assigned students at random to either the new method (this is the treatment group (bottom), $n=21$) or the traditional teaching method (this is the control group (top), $n=23$). The data from the study are reported in the graphs below.

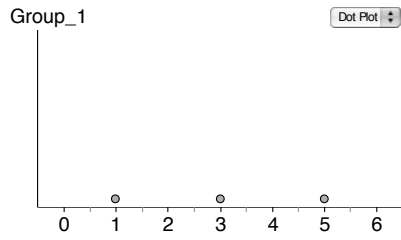
What evidence can you find in these data to determine whether the new method is better than the traditional method? Describe how you would go about making your decision and what, if any, reservations you might have about your conclusion.



(adapted from *Introduction to the Practice of Statistics*, Chapter 14, pp. 14-47 – 14-48)

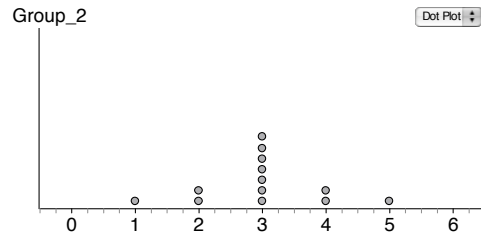
For each pair, which of the distributions below might be considered to have more variability? Explain your choice.

10.



A

Number of Quarters in Pocket

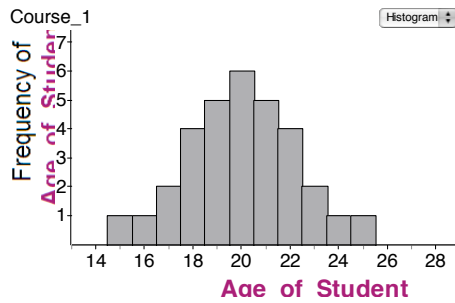


B

Number of Quarters in Pocket

Choice & explanation:

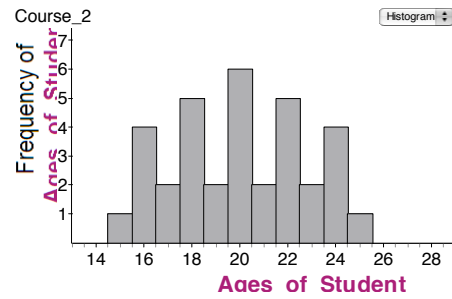
11. CPR Course 1



A

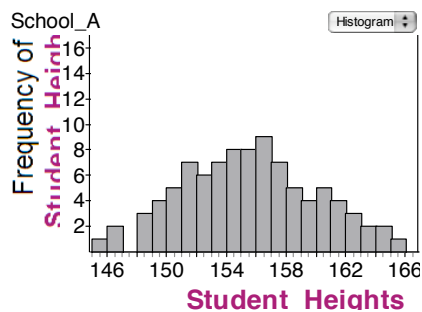
Choice & explanation:

CPR Course 2



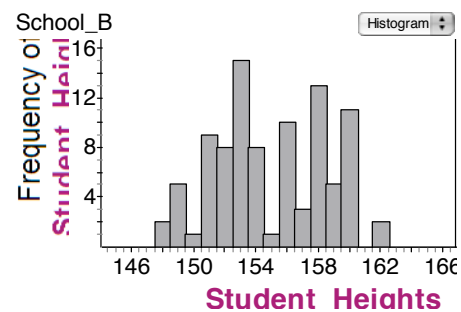
B

12.



A

Choice & explanation:



B

13.

The May 2006 issue of National Geographic contained the article “Misery for All Seasons—Allergies: A Modern Epidemic.” The following statement was made in the article: **“People who live with farm animals almost never have allergies”** (p. 127). The claim was made in reference to a theory called the hygiene hypothesis.

Seizing the opportunity to assess your students’ understanding of characteristics of well-designed studies, you decide to ask them to design a study that would help to provide evidence as to the truth of the claim. **Your task is not to design a study but rather to describe what you would expect a “high quality” student response to look like in this case. Please include the statistical ideas that are important to consider in this situation.** (e.g., what needs to be included and addressed to be considered thorough—no need to mention neatness or organization for presentation). *Please provide a description of the intended grade level and class to which you are referring. Your answer may differ by grade level or type of student.*

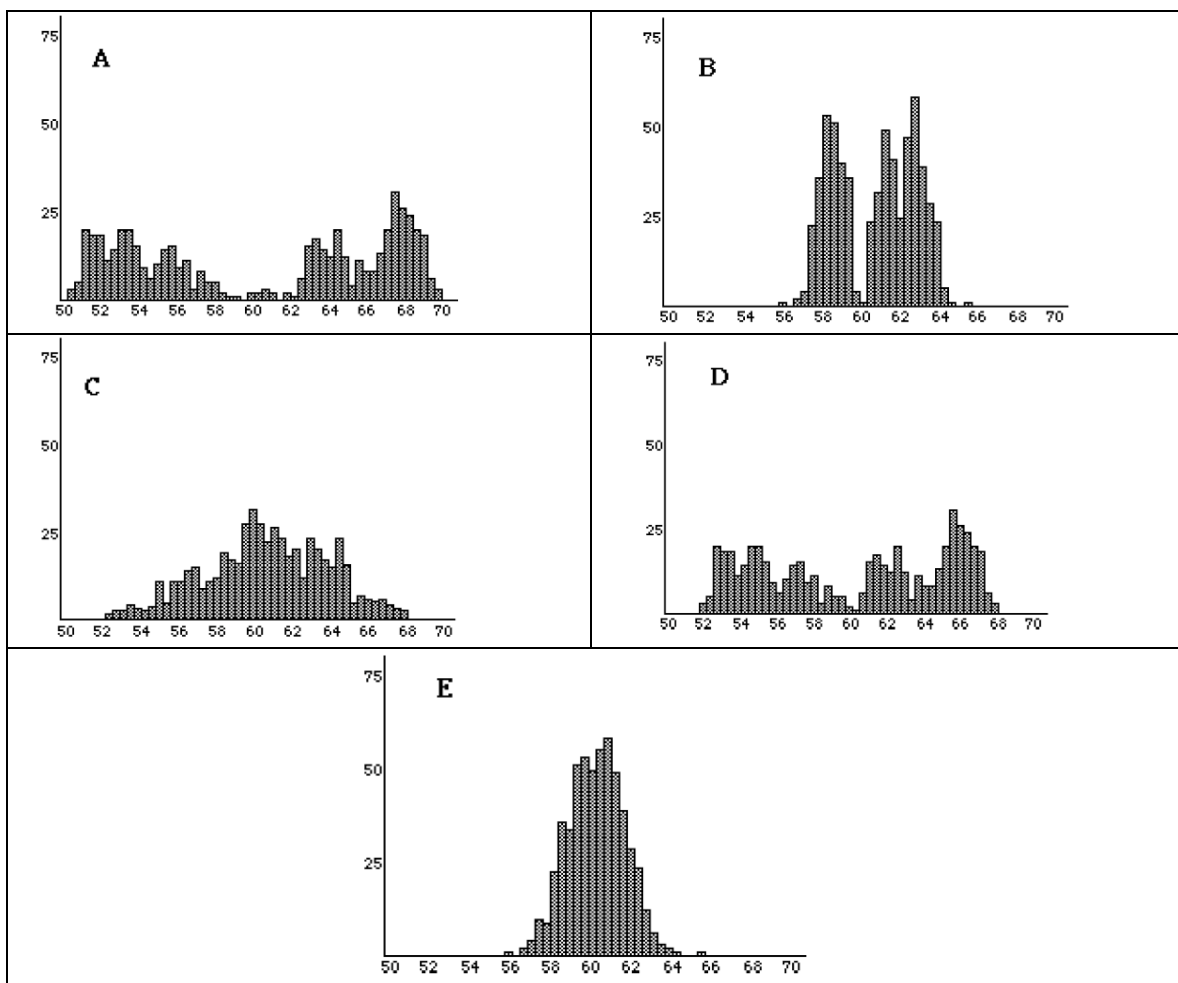
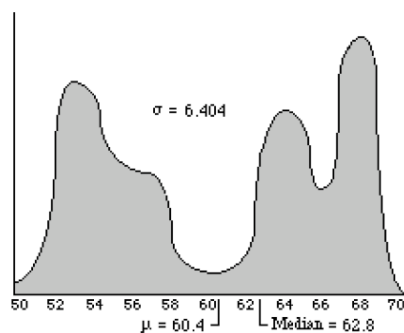
Intended Grade Level _____ Class _____

Description of class:

Characteristics of desired high quality student work:

Questions 14 – 20 on the next page refer to the graphs below. The first graph, labeled POPULATION, is the distribution of a *population* of test scores. Also shown are the population mean, median, and standard deviation. Each of the other five graphs labeled A to E represents a possible distribution of *sample means* for random *samples* of **size 25 or 4** drawn from the population.

POPULATION



For Questions 14 – 16, assume 500 samples of size 25 are randomly drawn from the POPULATION distribution shown on the previous page. For each question, circle the best response.

14. Which graph best represents the distribution of *sample means*.

A B C D E

15. I would expect the sampling distribution to be shaped more like:

- (A) a normal distribution
- (B) the population
- (C) can't tell

16. Which phrase comes closest to correctly completing the following sentence? I expect the sampling distribution to have...

- (A) less variability than the population.
- (B) the same variability as the population.
- (C) more variability than the population.
- (D) can't tell

For Questions 17 – 20, assume 500 samples of size 4 are randomly drawn from the POPULATION distribution shown on the previous page. For each question, circle the best response.

17. Which graph best represents the distribution of *sample means*.

A B C D E

18. I would expect the sampling distribution to be shaped more like:

- (A) a normal distribution
- (B) the population
- (C) can't tell

19. Which phrase comes closest to correctly completing the following sentence? I expect the sampling distribution to have...

- (A) less variability than the population.
- (B) the same variability as the population.
- (C) more variability than the population.
- (D) can't tell

20. Which phrase comes closest to correctly completing the following sentence? I expect the sampling distribution referred to in Question 14 to have...

- (A) less variability than the sampling distribution I chose in Question 17.
- (B) the same variability as the sampling distribution I chose in Question 17.
- (C) more variability than the sampling distribution I chose in Question 17.
- (D) can't tell

Appendix B
Post-Assessment

☐

Check this box if you
do **NOT** agree to
allow use of your
responses for
research purposes

Code # _____

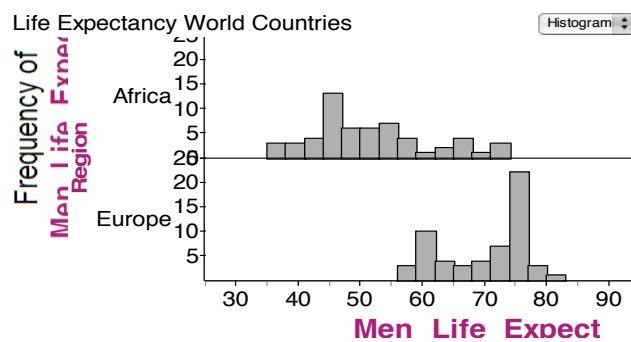
High School Mathematics Teachers' Evolving Understanding of Comparing Distributions

Summer 2006

Statistics Comfort-level. Please rate your level of comfort with each topic listed below by circling the level that best corresponds to a rating with **1 being very low/none** and **5 being high comfort**:

	low comfort		med comfort		high comfort
Descriptive statistics (mean, standard deviation, z-score)	1	2	3	4	5
Statistical Graphs (histogram, boxplot, bar graph)	1	2	3	4	5
Distributions (normal, chi-square, probability density functions)	1	2	3	4	5
Experimental Design (surveys, blocking, bias, sampling methods)	1	2	3	4	5
Correlation and Regression (least squares, r^2 , residuals, outliers)	1	2	3	4	5
Sampling Distributions (Central Limit Theorem)	1	2	3	4	5
Statistical Inference (t-tests, confidence intervals, chi-square tests, power, Type II error, ANOVA)	1	2	3	4	5

1. The following distributions represent **average life expectancies for men** from Africa and Europe. These life expectancies are computed for various regions within each country.



(Source: *Fathom2*; Taken from www.overpopulation.com which cites McDevitt, Thomas M., World population profile: 1998. Washington, DC: U.S. Government Printing Office, 1999)

How are the distributions of life expectancies for African men and European men similar?

How are the distributions of life expectancies for African men and European men different?

What can you say about the life expectancies of men from Africa and Europe? Please be specific.

2. Given the average winter temperature in cities X and Y, explain briefly how you would decide which of the following two events is more unusual: a 5 degree winter day in city X or a 5 degree winter day in city Y.

3. American males must register at a local post office when they turn 18. In addition to other information, the height of each male is obtained. The national average height for 18-year-old males is 69 inches (5 ft. 9 in.). Every day for one year, 5 men registered at a small post office and about 50 men registered at a large post office. At the end of each day, a clerk at each post office computed and recorded the average height of the men who registered there that day.

Which of the following predictions would you make regarding the number of days on which the average height for the day was more than 71 inches (5 ft. 11 in.)?

- a. The number of days on which the average heights were over 71 inches would be greater for the small post office than for the large post office.
- b. The number of days on which the average heights were over 71 inches would be greater for the large post office than for the small post office.
- c. There is no basis for predicting which post office would have the greater number of days.

Explain your choice and feel free to include sketches in your explanation.

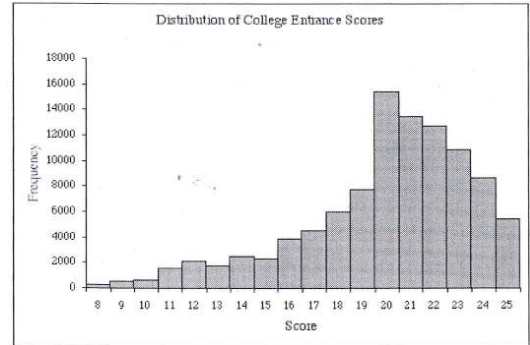
4. The **graph below right** shows the distribution of high school students' scores on a particular college entrance exam. Circle TRUE, FALSE, or CAN'T TELL **for each of the following statements.** (Graph from Ben-Zvi & Garfied (2004), p. 313)

- a. The distribution is skewed right. TRUE FALSE CAN'T TELL
- b. The median is greater than the mean. TRUE FALSE CAN'T TELL

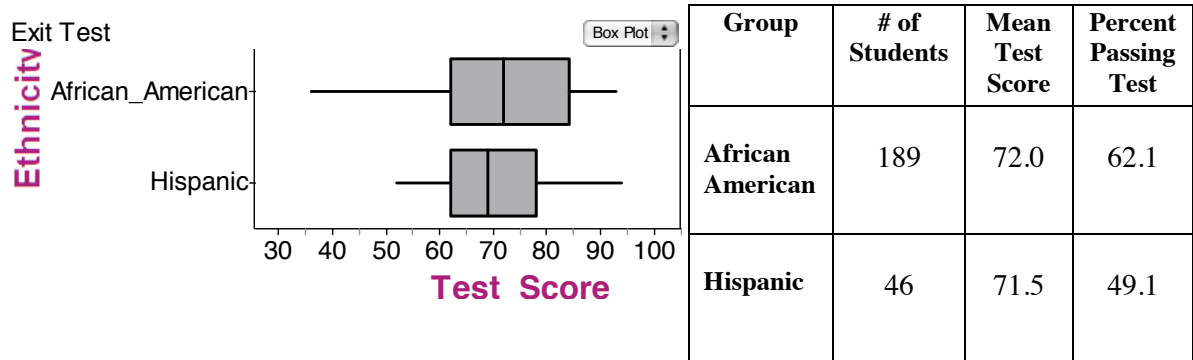
5. In the graph at right, how many students' scores are above the mean?

Choose one:

- A) Exactly half of the students' scores are above the mean.
- B) More than half of the students' scores are above the mean.
- C) Less than half of the students' scores are above the mean.
- D) I cannot answer the question without calculating the proportion of schools above the mean.



6. The pair of boxplots below represents the performance of two groups of 11th grade students from an urban high school in Louisiana on a 2005 district-mandated test. The top boxplot describes the performance of 189 African-American students while the bottom boxplot represents the performance of the 46 Hispanic students in the school. For reporting purposes, the class is considered “low-performing” if less than 50% of the students in any subgroup pass the exam. A score of 70 is considered passing. Additional information is provided in the table.



List three conclusions that would complete the following sentence: “By comparing the performance of Hispanic students with the performance of African-American students, I would draw the following conclusions...”

- 1)
- 2)
- 3)

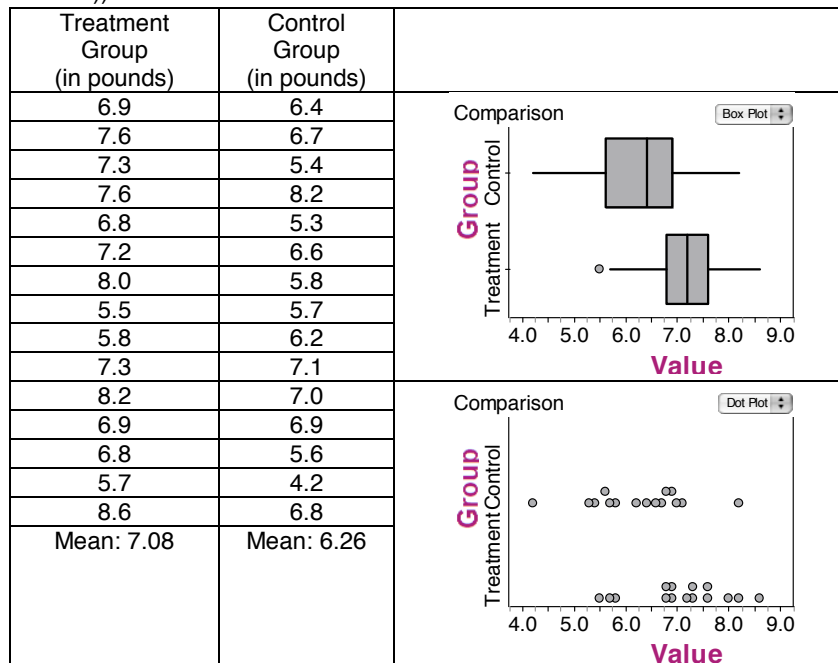
7. When a set of data has suspect outliers, which of the following are preferred measures of center and of variability for describing the distribution?

- A) mean and variance
- B) mean and standard deviation
- C) median and interquartile range
- D) median and range
- E) mean and range
- F) can't tell

8.

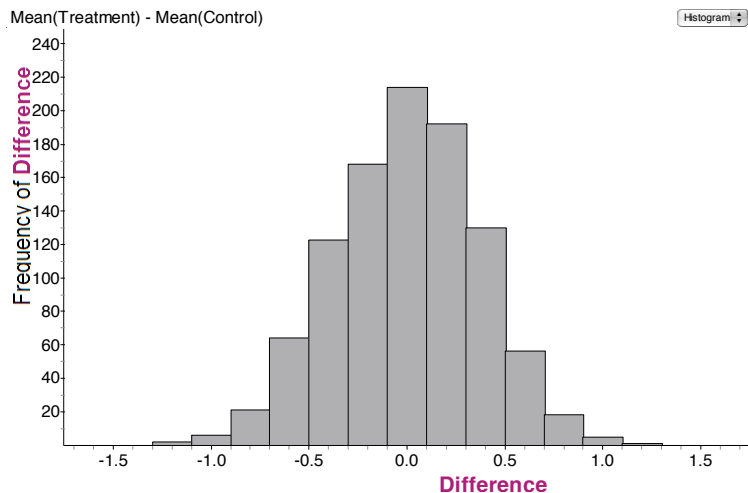
Is there a difference between treatments to prevent low birthweights? A research study was conducted to determine whether the use of a drug A prevents low birthweights (Rosner, 1982, p. 257). The study included 30 women assigned at random to two groups of size 15. The treatment group received doses of drug A while the control group did not. The birthweight data, accompanying boxplots and dotplots for the treatment and control groups are shown in the table below.

(Source: <http://www.resample.com/content/text/22-Chap-18.pdf> Simon, J.L. (1997). Resampling: The New Statistics (2nd Edition)).



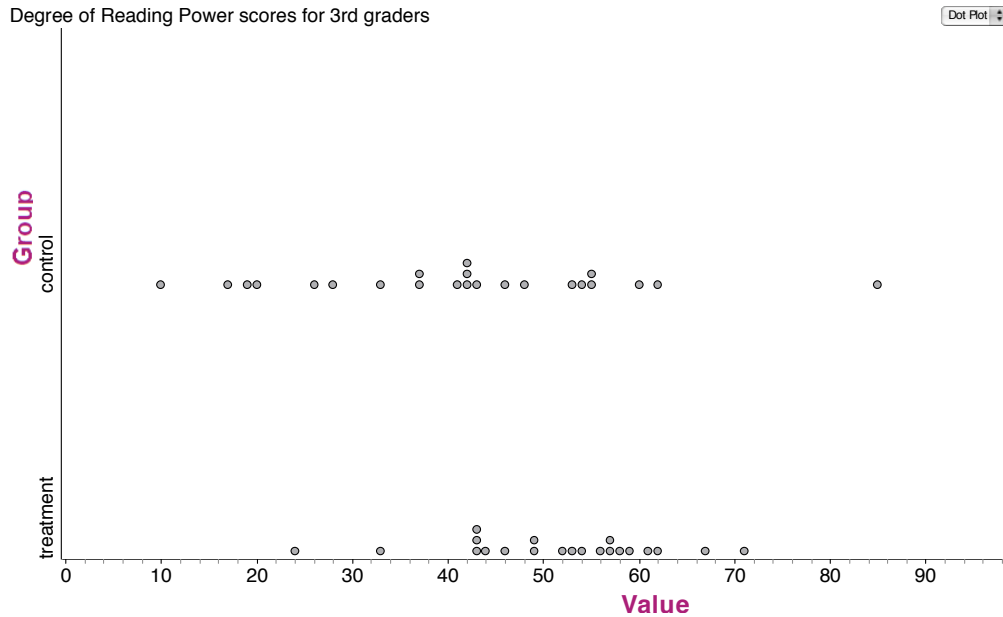
A randomization test was conducted to help determine whether the difference in mean birthweight between the treatment and control groups is statistically significant. The randomization distribution is below. **What do you conclude? Please explain.**

Randomization Distribution



9. A study was conducted to determine whether new “directed reading activities” improve the reading ability of elementary school students, as measured by their Degree of Reading Power (DRP) scores. The study assigned students at random to either the new method (this is the treatment group (bottom), $n=21$) or the traditional teaching method (this is the control group (top), $n=23$). The data from the study are reported in the graphs below.

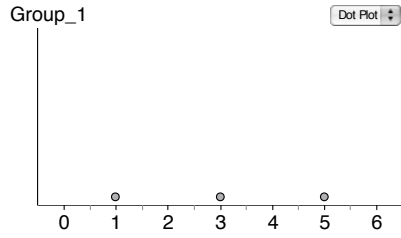
What evidence can you find in these data to determine whether the new method is better than the traditional method? Describe how you would go about making your decision and what, if any, reservations you might have about your conclusion.



(adapted from *Introduction to the Practice of Statistics*, Chapter 14, pp. 14-47 – 14-48)

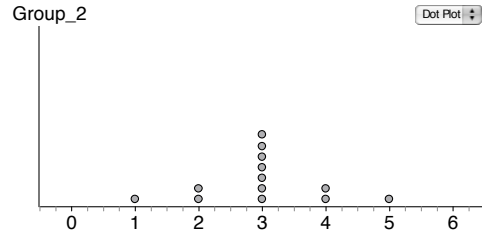
For each pair, which of the distributions below might be considered to have more variability? Explain your choice.

10.



A

Number of Quarters in Pocket

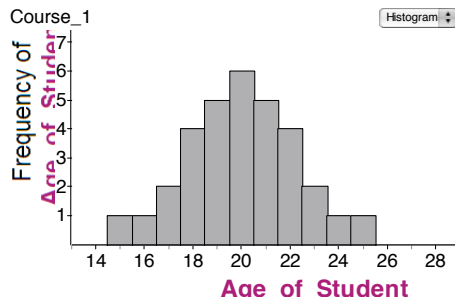


B

Number of Quarters in Pocket

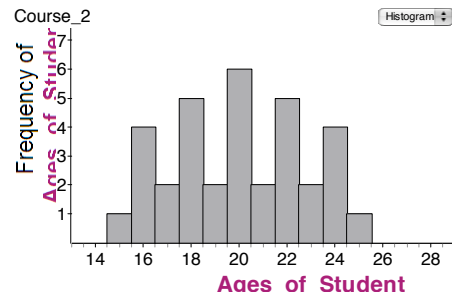
Choice & explanation:

11. CPR Course 1



Age of Student

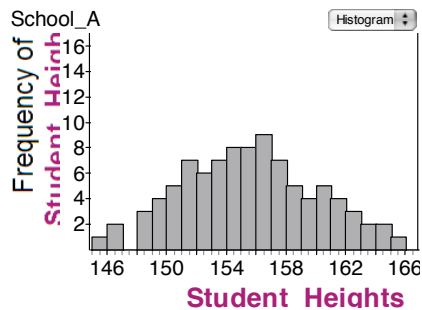
CPR Course 2



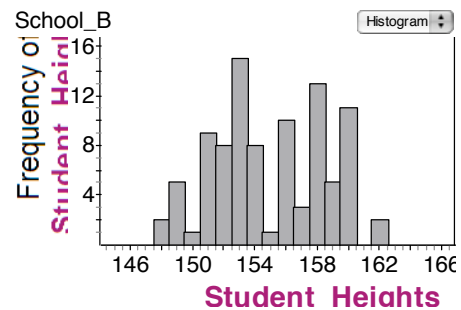
Ages of Student

Choice & explanation:

12.



Student Heights



Student Heights

Choice & explanation:

13.

According to a report in USA Today, the percentage of children 6 to 17 who are overweight has tripled over the past two and a half decades. Edward Sondik, director of the National Center for Health Statistics, suggests this trend is setting children up to be heavy adults and prone to such illnesses as diabetes, high blood pressure and heart disease. The *American's Children in Brief: Key National Indicators of Well-Being, 2006* report found, 18% of children ages 6 to 17 were overweight.

(<http://www.healthscout.com/news/68/533814/main.html>, retrieved 7/26/06).

Seizing the opportunity to assess your students' understanding of characteristics of well-designed studies, you decide to ask them to design a study that would help to provide evidence as to the truth of the claim. **Your task is not to design a study but rather to describe what you would expect a "high quality" student response to look like in this case. Please include the statistical ideas that are important to consider in this situation.** (e.g., what needs to be included and addressed to be considered thorough—no need to mention neatness or organization for presentation). *Please provide a description of the intended grade level and class to which you are referring. Your answer may differ by grade level or type of student.*

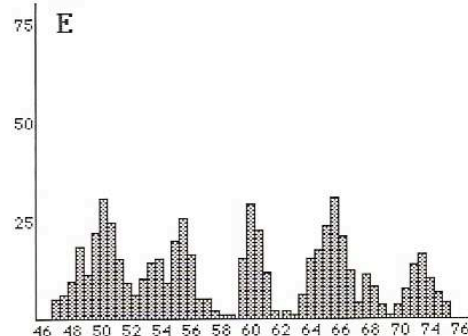
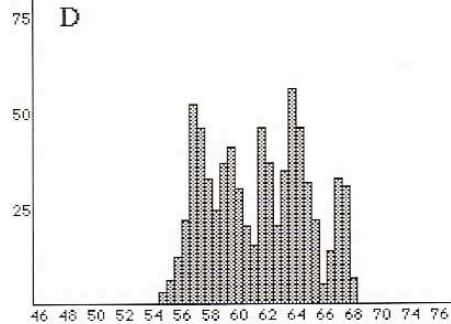
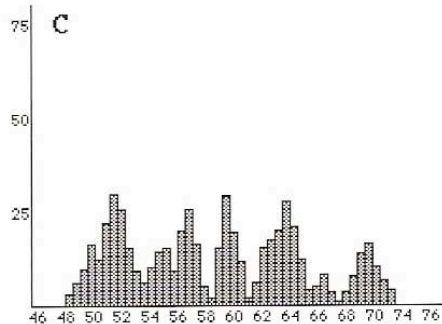
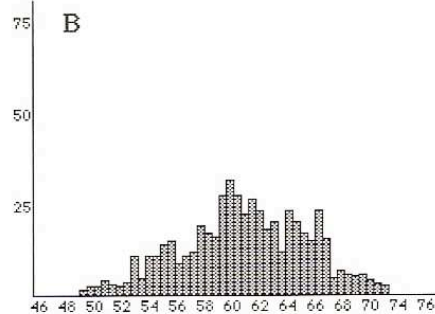
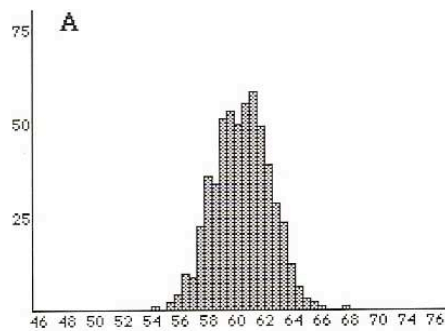
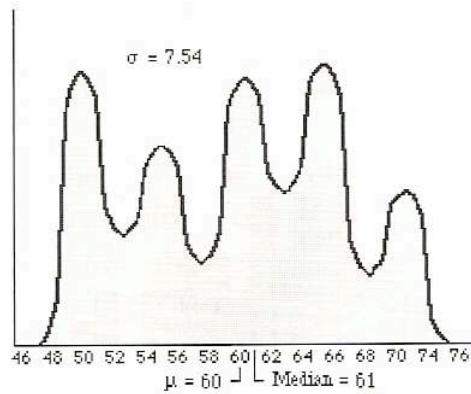
Intended Grade Level _____ Class _____

Description of class:

Characteristics of desired high quality student work:

Questions 14 – 20 on the next page refer to the graphs below. The first graph, labeled **POPULATION**, is the distribution of a *population* of test scores. Also shown are the population mean, median, and standard deviation. Each of the other five graphs labeled **A** to **E** represents a possible distribution of *sample means* for random *samples* of **size 16 or 4** drawn from *the* population.

POPULATION



For Questions 14 – 16, assume 500 samples of size 16 are randomly drawn from the POPULATION distribution shown on the previous page. For each question, circle the best response.

14. Which graph best represents the distribution of *sample means*.

A B C D E

15. I would expect the sampling distribution to be shaped more like:

- (A) a normal distribution
- (B) the population
- (C) can't tell

16. Which phrase comes closest to correctly completing the following sentence? I expect the sampling distribution to have...

- (A) less variability than the population.
- (B) the same variability as the population.
- (C) more variability than the population.
- (D) can't tell

For Questions 17 – 20, assume 500 samples of size 4 are randomly drawn from the POPULATION distribution shown on the previous page. For each question, circle the best response.

17. Which graph best represents the distribution of *sample means*.

A B C D E

18. I would expect the sampling distribution to be shaped more like:

- (A) a normal distribution
- (B) the population
- (C) can't tell

19. Which phrase comes closest to correctly completing the following sentence? I expect the sampling distribution to have...

- (A) less variability than the population.
- (B) the same variability as the population.
- (C) more variability than the population.
- (D) can't tell

20. Which phrase comes closest to correctly completing the following sentence? I expect the sampling distribution referred to in Question 14 to have...

- (A) less variability than the sampling distribution I chose in Question 17.
- (B) the same variability as the sampling distribution I chose in Question 17.
- (C) more variability than the sampling distribution I chose in Question 17.
- (D) can't tell

Appendix C

Pre-Intervention Interview Protocol

Potential Pre-Intervention Interview Questions
PROTOCOL
(anticipated time: 20-30 minutes)

BACKGROUND QUESTIONS

- **On the pre-assessment, you indicated _____ as your background in statistics. Describe what you remember of what you learned and a little about your experience with statistics. Are there reasons why you remember those particular topics? Do you recall topics that were particularly troublesome to learn? Please explain.**
- **Please describe any statistical topics or ideas you think all students should understand before leaving high school?**

STATISTICS QUESTIONS

I.

What do you look for (or to which characteristics do you refer) when comparing two distributions presented as

- **box plots**
- **line plots**
- **histograms**

Refer to the boxplot (#6), dotplot (#9), and histogram (#1) questions from the pre-assessment for concrete examples and possible follow-up

Do you have a preference in graphical representations of data? Please elaborate. [When looking at a concrete example, what particular aspects of the graph do you reason from? How do you make sense of distributions in these different representations? Is there information that is gained or lost when you have only one of the representations? . . .]

II.

Describe what variability means to you. [Look at pre-assessment questions 10-12 for clarification; may pursue relative to range, standard deviation, bumpiness. . . .]

III.

When data is used to support statements like “In a given situation/context. . . , the difference in means for the two groups is significantly different.” What does “significantly different” mean to you? [May use question #9 to refer to—would you consider the treatment group to have performed significantly better than the control group? Why/why not? Would additional information be helpful? Like what?]

IV.

Question 13 asked you to describe characteristics of a high quality student paper that would Here’s your response to this question (show paper). [May follow-up on specifics based upon individual solutions]. **Provide an example or two of researchable questions that your students might generate for this context. Describe how their research question might shape the design of their study. How would this impact the overall quality of their paper.** [Would you expect your students to construct an observational study or an experiment and how would that choice impact what they may plan?]

V.

Suppose that on a hiring examination, 35 out of 40 women pass and 20 out of 40 men pass. Under the assumption that men and women are equally likely to pass, a simulation was performed 500 times. The table below shows the number of men who passed. What can you conclude from this information?

[will have handout for participant; may have a histogram of the distribution just in case interpreting the tabular data is problematic]

(Adapted from Activity Based Statistics (Schaeffer, et. al., 2004))

Number of Males (out of 40) Who Passed	Frequency
19	3
20	4
21	6
22	6
23	18
24	29
25	47
26	49
27	69
28	76
29	62
30	59
31	36
32	15
33	11
34	7
35	2
36	1

VI.

OPTIONAL

(provided there exists ample time and reason to pursue additional questions of interest on the pre-assessment)

e.g., Look back at Question #9. Consider some possible reservations that a reader/researcher might have relative to the context of the question?

Appendix D

Post-Intervention Interview Protocol

Potential Post-Intervention Interview Questions
PROTOCOL
(anticipated time: approximately 45 minutes)

Please describe any statistical topics or ideas you think all students should understand before leaving high school?

STATISTICS QUESTIONS

I.

SEE PARTICIPANT HANDOUT (last page). Need a comparison question that supports the comparison of distributions both with and without the use of technology (AIDS DATA)

Looking for evidence of choice of technology (*CPMP-Tools*, Fathom, graphing calculator), choice of representations, and general argument. What understanding of comparing distributions is apparent?

II.

What do you look for (or to which characteristics do you refer) when comparing two distributions presented as

- **box plots**
- **line plots**
- **histograms**

Use #I. as a context for this question. If necessary, refer to the boxplot (#6), dotplot (#9), and histogram (#1) questions from the pre-assessment for concrete examples and possible follow-up

Do you have a preference in graphical representations of data? Please elaborate. [When looking at a concrete example, what particular aspects of the graph do you reason from? How do you make sense of distributions in these different representations? Is there information that is gained or lost when you have only one of the representations? **Has your ability to interpret distributional data in these forms changed since the M2RI session? In what ways?**]

III.

Describe what variability means to you. [Look at pre-assessment questions 10-12 for clarification; may pursue relative to range, standard deviation, bumpiness. . . . **Describe how the concept of variability has changed as a result of the professional development program.**]

IV.

When data is used to support statements like “In a given situation/context. . . , the difference in means for the two groups is significantly different.” What does “significantly different” mean to you? [May use question #9 to refer to—would you consider the treatment group to have performed significantly better than the control group? Why/why not? Would additional information be helpful? Like what? **NEW: Look at question #8 for the randomization test interpretation—how is significantly different interpreted? In what ways does the randomization test support making comparisons between distributions?**]

V. Look carefully at the responses for questions 14-20. Ask teachers to try to articulate their understanding of the task. If necessary remind them of the sampling distribution task and modeling activity that preceded the central limit theorem discussion in order to help reason through the problem.

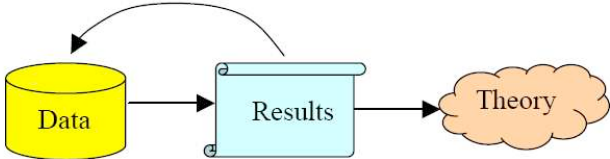
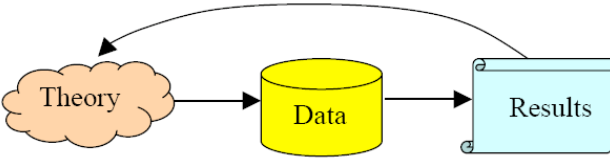
VI.

The following articles and/or chapters were recommended for reading during and/or following the M²RI professional development program. Please indicate the articles that you have read by the time of this interview by checking the appropriate box. Additionally, place a check by the row if you read the article prior to the post-test.

☐ Yes ☐ No	The Structure of Fathom, pp. 9-15 (from Fathom Learning Guide—Key Curriculum Press, 2005)
☐ Yes ☐ No	Levasseur, K., & Cuoco, A. (2003). <i>Mathematical habits of mind</i> . In H. L. Schoen (Ed.), Teaching mathematics through problem solving: Grades 6-12 (pp. 27-38). Reston, VA: NCTM.
☐ Yes ☐ No	Barbella, P., Denby, L., & Landwehr, J. (1990). <i>Beyond exploratory data analysis: The randomization test</i> . The Mathematics Teacher, 83(February), 144-149.
☐ Yes ☐ No	Rossman, A., Chance, B., & Medina, E. (2006). <i>Some important comparisons between statistics and mathematics, and why teachers should care</i> . National Council of Teachers of Mathematics Sixty-Eighth Yearbook, Chapter 22.
☐ Yes ☐ No	USA Today, March 2, 2006— <i>To head off allergies, expose your kids to pets and dirt early. Really.</i>
☐ Yes ☐ No	Franklin, C. & Garfield, J. (2006). <i>The GAISE Project: Developing statistics education guidelines for grades pre-K-12 and college courses</i> . National Council of Teachers of Mathematics Sixty-Eighth Yearbook, Chapter 24.
☐ Yes ☐ No	Shaughnessy, J. M. (2006). <i>Research on students' understanding of some big concepts in statistics</i> . National Council of Teachers of Mathematics Sixty-Eighth Yearbook, Chapter 7.
☐ Yes ☐ No	OPTIONAL for A.P. Stats Teachers: Bock, D. & Velleman, P. (2006). <i>Why variances add—and why it matters</i> . National Council of Teachers of Mathematics Sixty-Eighth Yearbook, Chapter 25.
☐ Yes ☐ No	OPTIONAL for A.P. Stats Teachers: Koehler, M. (2006). <i>Using graphing calculator simulations in teaching statistics</i> . National Council of Teachers of Mathematics Sixty-Eighth Yearbook, Chapter 18.
☐ Yes ☐ No	Grouws, D. (2003). <i>The teacher's role in teaching mathematics through problem solving</i> . In H. L. Schoen (Ed.), Teaching mathematics through problem solving: Grades 6-12 (pp. 129-142). Reston, VA: NCTM.
☐ Yes ☐ No	Rubin, A. & Hammerman, J. (2006). <i>Understanding data through new software representations</i> . National Council of Teachers of Mathematics Sixty-Eighth Yearbook, Chapter 17.
☐ Yes ☐ No	Groth, R. (2006). <i>Engaging students in authentic data analysis</i> . National Council of Teachers of Mathematics Sixty-Eighth Yearbook, Chapter 4.
☐ Yes ☐ No	Copes, L. & Shager, N. K. (2003). <i>Phasing problem-based teaching into a traditional education environment</i> . In H. L. Schoen (Ed.), Teaching mathematics through problem solving: Grades 6-12 (pp. 195-206). Reston, VA: NCTM.
☐ Yes ☐ No	Stein, M. K. & Boaler, J. (2003). <i>Teaching mathematics through problem solving: Research perspectives</i> . In H. L. Schoen (Ed.), Teaching mathematics through problem solving: Grades 6-12 (pp. 245-256). Reston, VA: NCTM.
☐ Yes ☐ No	Rasmussen, C., Yackel, E., King, K. (2003). <i>Social and sociomathematical norms in the mathematics classroom</i> . In H. L. Schoen (Ed.), Teaching mathematics through problem solving: Grades 6-12 (pp. 143-154). Reston, VA: NCTM.
☐ Yes ☐ No	Konold, C., & Higgins, T. L. (2003). <i>Reasoning about data</i> . In J. Kilpatrick, W. G. Martin & D. Schifter (Eds.), A Research Companion to Principles and Standards for School Mathematics. Reston, VA: NCTM.

VII.

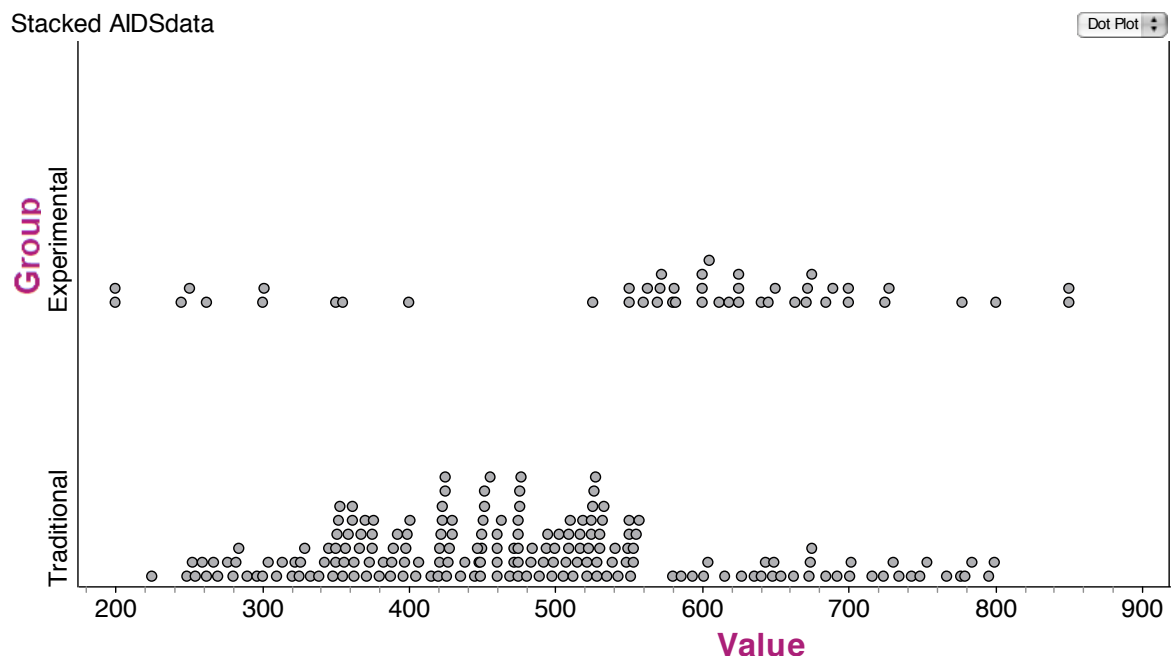
Makar (2004) studied teachers' behaviors while learning to use technology (more specifically Fathom) to conduct investigations. She categorized them as follows:

Check the category that best represents your view of how you use technology to conduct investigations	Category	Description of Category
	Wanderers	Use data to look for a theory. Spend a good deal of time “wandering” through various analyses that are not necessarily directly connected to a conjecture, but hopeful that something will jump out that can be tied back to a conjecture.  <pre> graph LR Data[(Data)] --> Results[Results] Results --> Theory((Theory)) Theory --> Data </pre>
	Wonderers	Lead by “I wonder” questions. Generally create a theory and then use the data to test their theory.  <pre> graph LR Theory((Theory)) --> Data[(Data)] Data --> Results[Results] Results --> Theory </pre>
	Answerers	Go into an investigation with a theory like the Wonderers, but without generating “I wonder” questions during investigation. Search for a particular piece of evidence to support or refute their conjecture and then directly state their conclusion.  <pre> graph LR Theory((Theory)) --> Data[(Data)] Data --> Results[Results] </pre>

I.
TASK:

The distributions below contain T-cell counts of two groups of AIDS patients who had enrolled in different treatment protocols. Compare these distributions. First, describe how you would begin to analyze these data without access to technology. Then demonstrate and discuss how, using your choice of technological tools, you would analyze these data in order to make a comparison.

$n_{\text{experimental}}=47$
 $n_{\text{treatment}}=187$



Appendix E

Written Reflection Prompts

Reflection #1

Code # _____

Based on today's experiences, please describe ways in which your understanding of some statistical ideas has changed (what do you understand better now than you did this morning—please explain?).

Reflection #2 Code #_____

Code # _____

- In your own words, describe what you have learned about the randomization test.

- What is it useful for?

- What does it help you to figure out?

- What questions do you have about the randomization test?

- On a scale from 1 to 10, rate the **ease of use** of *CPMP-Tools* to conduct the randomization test. (1 is low; 10 is high)

Circle one:

1	2	3	4	5	6	7	8	9	10
<i>difficult</i>				<i>medium difficulty</i>					<i>easy</i>

- On a scale from 1 to 10, rate your **current understanding** of the randomization test.

Circle one: 1 2 3 4 5 6 7 8 9 10
poor decent excellent

Reflection #3

Code # _____

- Describe insights or new ideas you have gained through the investigations so far regarding:

Measures of Center

Variability

Bias

Design of experiments

- What questions do you have about anything we have been doing so far?

- On a scale from 1 to 10, **rate your current feelings** about how your learning of *Fathom2* is going (1 is low; 10 is high)

Circle one: 1 2 3 4 5 6 7 8 9 10
difficult/frustrating *I'm getting there* *no problem*

- On a scale from 1 to 10, rate your **current understanding of the randomization test**.

Circle one: 1 2 3 4 5 6 7 8 9 10
poor *decent* *excellent*

Reflection #4 Code # _____

Code # _____

- On a scale from 1 to 10, rate your **current understanding of the randomization test**.

Circle one: 1 2 3 4 5 6 7 8 9 10
poor decent excellent

- What does “p-value” mean to you?

- What does “significantly different” mean to you?

Reflection #5

Code # _____

- Describe what you think is meant by the phrase “the difference is statistically significant.”
-
- When you think about comparing distributions, how have your ideas changed or grown this week—please describe.
-
- Describe your understanding of the relationship between the size of a sample and the related sampling distribution that can be generated.
-
- What questions do you still have about the randomization test?
-
- On a scale from 1 to 10, rate your **current understanding of the randomization test**.
- Circle one: 1 2 3 4 5 6 7 8 9 10
poor decent excellent
- On a scale from 1 to 10, rate your **current understanding of the relationship(s) between the size of a sample and the sampling distribution of the mean that can be generated from samples**.
- Circle one: 1 2 3 4 5 6 7 8 9 10
poor decent excellent
- On a scale from 1 to 10, **rate your current feelings** about how your learning of Fathom2 is going (1 is low; 10 is high)
- Circle one: 1 2 3 4 5 6 7 8 9 10
difficult/frustrating I'm getting there no problem

Appendix F

Assessment Scoring Rubrics

Scoring Rubrics for Pre- and Post-Assessments

Item #1—Q1, Comparing distributions: Histograms

NOTES: Since no information was provided in this item that would have included population of each region, it would be difficult (impossible), to accurately predict the average life expectancy for each country. Very few teachers took this into account and that may be a limitation of this question. However, if we treat the life expectancies as coming from equally sized regions, then direct comparison is fine.

REGIONS: This is a question of where one lived. Population density suggests these are means of means.

Task may be too sophisticated for the intended purpose—scoring and interpretation were handled by treating regions as equally weighted regions.

Score	General Criteria
4	- Addresses shape, center, and variability accurately by estimating these statistics either graphically or numerically - Uses statistical language appropriately Conclusion is in context with correct interpretation (does not use language of # of E vs. # of A)—Needs to notice that
3	- Demonstrates broad understanding of the question and major concepts necessary to respond—center, shape, and variability - All parts are correct and a correct answer is achieved - Conclusion may be interpreted in terms of # of E vs. # of A (common misinterpretation in this problem) Uses statistical language Some mention of variability—to include range-- The concept of statistical tendency becomes part of the discussion and conclusion about data
2	- Solution is not complete—addresses some but not all of the aspects of the problem - May use “normal” or “bell-shaped” language to describe shape and not refer to any skewness - Statistical language is limited - May not estimate the center or range of the distributions - Mentions mean, median, OR mode - May discuss range or <i>SD</i> (not necessary) but not beyond the descriptive level - May compare # of Africans to # of Europeans (not average) and thus misinterprets
1	- States obvious things from the pictures without evidence of statistical understanding or misinterprets graph - May lack statistical language - Inappropriate conclusions are drawn or overly general conclusions are suggested
0	blank

Item #2—Q2, Comparing distributions: Limitations of mean only (need distribution/variability)

This question requires some attention to typical temperatures as well as something about variability. (Makar—pretest 17% average & variation; posttest 69% average & variation)

Score	General Criteria
4	• Standard deviation of temperatures for both cities and determine how many standard deviations 90 degrees (5 degrees) is from the mean • OR • Used some idea of variability in the answer to differentiate from mean only • may provide an example that illustrates that variability is a factor AND uses the example to provide a full explanation
3	• range of average temperatures • location • OR uses standard deviation, but the explanation is somewhat problematic (talks about p-value) • may provide an example that illustrates that variability is a factor but may not provide a full explanation • recognizes the need for more information about distribution or cyclic nature of temperatures
2	• used average only, but provided example to help. Example might be problematic, but it is appropriate • may use more than 1 example to show the relationship between 2 average temps may vary (5—x---y; or x---5---y) • may confuse the randomization test idea using differences from 5 degrees • May recognize that averages are insufficient but does not suggest a viable alternative • May mention outliers throwing off average suggesting a distributional idea
1	• average only OR find out how many 90 degree (5 degree) days each city has. • argues from 1 specific case
0	blank

Item #3—Q3, Small sample variability

Because pretest did NOT require an explanation, scores for this item were 4 if correct, 1 if answered incorrectly, and 0 if left blank.

The scoring guide below was used for the qualitative analysis of the post-assessment only.

Score	General Criteria APPLIED ONLY TO POST-ASSESSMENT QUALITATIVE ANALYSIS
4	• explicitly recognizes that increased sample size corresponds to a sampling distribution with smaller standard error and sample means further away from the mean of the sampling distribution become more unusual (not necessary to use this language) • may understand the concept of sampling distributions (CLT)
3	• realizes that small samples may have unusual sample means, but explanation may be incomplete or problematic
2	• Pick correct answer and says “sample size” but the reasoning is not clear • may pick incorrect answer but explanation suggests they understand that small sample means are more variable (this helps to overcome the incorrect choice)
1	• recognizes the need for a randomization test to confirm, but does not seem to understand the variability vs. sample size concept • chooses wrong answer and explanation indicates misconception
0	blank

Item #4—Q4a, 4b, 5, 7—Characteristics of distributions—skewed, mean vs. median, outliers and variability

Score	General Criteria
4	• All 4 correct
3	• 3 of 4 correct
2	• 2 of 4 correct
1	• 1 of 4 correct
0	• 0 of 4 correct or blank

Item #5—Q6, Comparing Distributions: Boxplots

Using Makar’s (2004) scoring scheme (below) whereby each part, 1, 2, and 3, is scored on a 5-level scale. The maximum 2 scores are added to constitute the final score on this question; hence there are 10 points possible. In this study, the final score is rescaled to a 4-level system to match the remaining questions through using the linear transformation of $y=1/2*x-1$ ($10 \rightarrow 4$; $8 \rightarrow 3$; $6 \rightarrow 2$; $4 \rightarrow 1$).

Response Level	Description	Number of responses	Categories included	Sample responses
0	No response	11		N/A
1	Not based on data	1		-“Both groups are receiving the same quality of instruction.”
2	Comparison directly from the table	37	Higher average or percent passing, vague comparison, statement of number of students	-“The percent of Hispanic students passing the test is much higher than that of the African-American students.” -“There are more Hispanic students than African-American students.”
3	Some interpretation used	29	Low-performing status, mean scores, equal number of students passing, mention of high/low scorers	-“Less than 50% of African-American students passed the exam causing the school to be considered low-performing.” -“Hispanics and African-Americans have similar mean TAAS scores.”
4	Suggests statistical skill	28	Comparison of medians, range, shape, effect of sample size or outliers	-“The range of scores for Hispanic students is larger than that of African-American students.” -“The fact that there were so few black students may influence their test scores.”
5	Suggests distributional-view of the data or awareness of variability	12	Mentions variability, distribution or partial distribution (e.g. quartile)	-“Because the population is smaller, there is less variability in scores.” -“There are much lower scores in the lower quartile for Hispanic students.”

NOTES: Misuse of boxplots (Q1 compared to Q3) scores a 3 on Makar’s scale instead of 4. A 3 would include weak boxplot—maybe no comparison.

Language of “significant difference” without variability or distribution will not score 5, instead score 4. To score 5 using quartiles, the comparison should be clear and accurate (to describe how partial distributions differ).

Item #6—Q8—Simulation/Randomization Distribution

Score	General Criteria	Pretest Specifics	Posttest Specifics
4	Thorough interpretation of the problem.	Computes $\bar{x}=15.6$ boxes; $s_x=4.615$ *Discusses small # trials *Mentions at least 6 boxes needed to be successful and the need for more trials and provides a reasonable estimate—<u>includes a range</u>	Difference in mean birthweights is significant Approximates OR infers p-value Compares mean birthweights, $\text{diff}=.82$ *NECESSARY for level 4-- Draws a conclusion in the context of the problem
3	Reasoning more stochastically than deterministically	Average & skepticism Provides rationale for answer Reasoning goes beyond simulation Doesn't consider small # trials Center may be a range (e.g. 13-17)	Computes difference in birthweights = .82 Uses the r.d. distribution or the location of .82 on the distribution to reason that the difference is significant, but does not discuss p-value Conclusion may not be linked to context Idea may appear correct, but language may be shaky
2	Computation and additional basic interpretation of simulation	Average & additional reasoning $15.6 \rightarrow 16$—rounding suggests context interpretation Reasoning based only on simulation	Computes difference in birthweights = .82 Reasons from the boxplots or the dotplots and comes to the conclusion that the difference is not significant. A distributional tendency is apparent though. May use the r.d., but the reasoning is problematic
1	Limited computation	Computes average 15.6 only OR uses the range only OR uses the minimum or maximum ONLY OR avoids the mathematics/statistics in the problem (non-trusting) did not answer the intended question OR Didn't make a data-based argument	Computes difference = .82, but does not correctly interpret the r.d. distribution OR Lacks good explanation (avoids the r.d. reasoning)
0	blank	No answer	

Item #7—Q9, Comparing Distributions: Dotplots

Score	General Criteria
4	• mean or median • outliers OR range OR standard deviation • Proportional argument • treatment>control • 90% (19/21) T>42 • 56% (13/23) C>42 • IQR • Reservations may include: 1. Small sample 2. Teacher effect 3. Lack of replication 4. Pretest scores—equivalence of groups 5. Other
3	• Uses mean/median/quartiles to argue BUT may not use specific statistics • Must attend to some reservations (may elude to statistics: treatment mean > control mean. . . • Attends to center AND variation in some way • Conclusion is based on data in context • May use additive argument (groups are similar size) • Reservations are given OR the suggestion of a randomization test is made
2	• Clustering—range • Conclusion is reasonably based on data • Needs to take a position • May have reasonable idea but does not provide evidence from the data • Reservations are not attended to • May suggest randomization distribution without detail • May discuss overlap and reach incorrect conclusion, but nothing else is estimated.
1	• Calculate nothing OR • Does not reference statistical measures OR • Conclusion is based on faulty reasoning
0	blank

Item #8—Q10, 11, 12—Comparing Distributions: Variability

Part 3 (#12) was from Makar's (2004) study. This question is intended to explore thinking about variability and whether individuals think of variability as bumpiness, range, or deviation from center or otherwise. (Makar indicated that reasoning was problematic; most listed range. Pretest 61% correct; Posttest 89% correct).

Score	General Criteria
4	Reasoning indicates variation as deviation from center and respect for range— chooses all 3 correctly or reasoning is CONSISTENT with deviation from a central anchor perspective.
3	Reasoning is good for 10,11, but regresses for 12 OR May get all 3 choices correct but reasoning does not CONSISTENTLY indicate deviations from a central anchor.
2	Tends to confuse normal distribution with variation May get #10, 11, 12 correct—suggests some understanding, but explanations are confused. May use a variety of strategies to select; Average distance from center is not a clear strategy May suggest range, but needs to apply it in #12 and do something reasonable for 10, 11.
1	Reasoning indicates limited or no idea of variation May get #12 right with range but exhibits confusion with 10, 11 when ranges match.
0	blank

Item #9—Q13—Experimental Design

Important ideas: 1) Formulate question; 2) Collect data; 3) Analyze data; 4) Draw conclusion. If an experiment is anticipated, then one would expect 1) random assignment; 2) control/experimental treatment; 3) sufficient number of subjects.

Depending on perspective, this may be an observational study or an experiment.

Score	General Criteria	Pretest Specifics	Additional Posttest Specifics
4	• Formulate • Collect • Analyze • Interpret	Interprets task reasonably Defined farm animals Some sort of comparison (control) Some attention to statistics Sample size Randomization *Needs research question (formulate question)	Either context specific or sufficiently generic detail to assume knowledge
3	• Collect • Analyze • Interpret	May lack research question Needs to address sample size; randomness, definitions, & comparison Needs to include context	
2	• Collect and one other	Explanation indicates too broad a perspective (can't really be done—not researchable) Lists criteria but not in a cohesive manner—BIG IDEAS OF DESIGN INCLUDED Does not deal with definitions (farm animals/allergies) Lacks connection to context	May discuss the 3 criteria for controlled experiment, but does not link to context
1	• Some attempt to answer the question	Discusses generic ideas w/o focus OR Provides list of criteria which is insufficient OR Answer is too vague or lacking specificity or direction	
0	blank		

Item #10—Q14, 15, 16, 17, 18, 19, 20—Sampling Distributions

Score	General Criteria
4	All 7 questions were correct. Because of the nature of the questions, and the fact that it is very unlikely to achieve correct answers by guessing.
3	Either 6 correct OR Correct answers for 14, 15, 16, (this suggests coordination for large sample size) and 20 (indicative of variability of sampling distribution decreasing with sample size) plus either a correct answer on #17 or a choice that suggests awareness of decreasing variability with sample size, plus either 18 or 19 correct. Other configurations of 5 correct were looked at on an individual basis to determine whether the combination of answers suggested a strong likelihood of understanding of sampling distributions
2	At least 4 correct, but not at the level required to score a 3. Deferring to the probability of 4 correct, it is unlikely that this would happen just by chance alone, but with 3 (or 2 in the case of 5 correct) incorrect, it becomes difficult to determine whether the incorrect answers are the result of guessing or of not knowing. 4 correct suggests at least a developing understanding of sampling distributions
1	Guessing cannot be ruled out. A score of 1 means that between 1 and 3 questions were answered correctly.
0	Blank or 0 questions answered correctly

Appendix G

Professional Development Facilitator Guide Pages

ACTIVITY 1: Statistics: Activity, Technology, & EDA

Overview:

Statistics is one of the mathematical strands for which teachers have the least preparation to teach. This session will introduce National and State-level grades 9-12 standards, benchmarks, and grade level content expectations for statistics. Additionally, it will serve as the kick-off for four days of activity-based, technology intensive, exploratory data analysis with sense-making at its core. The randomization test will be introduced.

Goals/Objectives:

Understand that statistical knowledge is appropriate and expected of all students upon high school graduation.

Experience a statistical investigation which embodies an activity-based, exploratory data analysis (EDA), sense-making perspective.

Develop a conceptual understanding of mean absolute deviation, standard deviation, and sample standard deviation.

Understand various measures of center and spread and be able to reason graphically and in context about the appropriateness of both.

Compare distributions from experimental data and begin to think about whether differences are attributed to chance or something else.

Materials to Conduct the Activity:

Notecards (two colors)

Fathom2 and *CPMP-Tools*

Transparencies:	T-1-2	NCTM Grades 9-12 Data Analysis & Probability Standards
	T-3	Matching Plots to Variables
	T-4	Variable of State--Boxplots
	T-5	Geometric Interpretation of Standard Deviation
	T-6	Orange Trees and Fertilizer
	T-7	Dividing by $n-1$ in Sample Standard Deviation
	T-8-9	Matching Statistics to Plots
	T-10	<i>CPMP-Tools</i>

Handouts:	H-1-11	State and National Recommendations for Statistics in the Secondary Curriculum
	H-12-13	Orbital Express—Part I
	H-14-15	Interpreting Box plots and Back-to-back Stemplots (from CPMP Course 1, Unit 6, Patterns in Data, p. 26 #24; p.

	38 # 4)
H-16-20	Matching Plots to Variables
H-21	Activity 6-3: Variables of State (from Rossman, Chance, & Lock—Workshop Statistics: Discovery with Data and Fathom, p. 131)
H-22-23	Geometric Interpretation of Standard Deviation (from Interactive Mathematics Program—Is There Really a Difference)
H-24-29	Hirsch, et al. (2006). <i>Patterns in Data</i> . Course 1, Unit 2, Measuring Variability: The Standard Deviation. Core-Plus Mathematics Project 2 nd Edition draft materials. pp. 45-74
H-30	Orange Tree/Fertilizer Problem—CPMP Unit 6, Course 1, Lesson 2, On Your Own Applications #9 p. 59
H-31	Why Do We Divide By $n - 1$ in the Formula for the Sample Standard Deviation?
H-32	Sample Statistics as Estimators
H-33-36	Matching Statistics to Plots
H-37-40	Orbital Express—Parts II & III
Readings:	R-1*** The Structure of Fathom, pp. 9-15 (from <i>Fathom2 Learning Guide</i> —Key Curriculum Press, 2005)
R-2*	Levasseur, K., & Cuoco, A. (2003). <i>Mathematical Habits of Mind</i> . In H. L. Schoen (Ed.), <i>Teaching mathematics through problem solving: Grades 6-12</i> (pp. 27-38). Reston, VA: NCTM.
R-3**	Barbella, P., Denby, L., & Landwehr, J. (1990). <i>Beyond Exploratory Data Analysis: The Randomization Test</i> . <i>The Mathematics Teacher</i> , 83(February), 144-149.
R-4	Hirsch, et al. (2006). Unit 2, Course 1. <i>Patterns in Data</i> . Core-Plus Mathematics Project 2 nd Edition draft materials. pp. 45-74
R-5	USA Today, March 2, 2006— <i>To head off allergies, expose your kids to pets and dirt early. Really.</i>
R-6***	Rossman, A., Chance, B., & Medina, E. (2006). <i>Some Important Comparisons between Statistics and Mathematics, and Why Teachers Should Care</i> . National Council of Teachers of Mathematics Sixty-Eighth Yearbook, Chapter 22.

*** Indicates a reading assignment required for the next day

** Indicates a desired reading assignment for the next day, but if time is short, may read later

* Great if you can read it before the next session, but not essential

Other readings are for your reference and you may read at your leisure :)

Suggested Teaching Procedures:

Activity 1.1: Welcome and Review of Professional Recommendations and Standards for Statistics (30 minutes)

Welcome everyone back and begin using T-1-2 to remind participants of the NCTM Principles & Standards. Only the high school expectations are included on the transparency, but the H-4 contains middle and high school standards and expectations. Take this time to engage in conversation about the role of statistics in the curriculum and allow participants to review the middle and high school content expectations as well as the State curriculum framework documents in their packet (H-1-11). This is the only time during this session where middle and high school teachers will be together, so try to make sure the conversation is across grade levels so that each grade band of teachers is cognizant of the expectations for statistical literacy, reasoning, and thinking at their level. Refer to this material often during the sessions.

Activity 1.2: Orbital Express Part I (1 hour)

You will need at least two kinds of paper for each group for this activity (copier paper and paper towel, or something else). Groups will need measuring tapes and/or yard sticks, two colors of notecards (cut into smaller pieces to record and later shuffle) Construct groups of size 4 using some random process and then assign roles for the groups. Facilitate a conversation about issues associated with data collection and try to encourage good quality control. Things that should come out in the introduction include: 1) How to maintain consistency in dropping, measuring, and recording data, and 2) Is it wise to do all of the drops of one vehicle and then all of the drops of the other?

Groups will construct a poster of their data using some representation of their choice. They need decide which vehicle is best and communicate what they mean by that. What is their conclusion? How confident are they in their results? As groups share their analysis and conclusions, be explicit with questions that will elicit their understanding and encourage participants to ask questions of each other (working on norms of careful listening, non-evaluative stance, using evidence to support statements, and equitable airtime as well as formative assessment relative to what constitutes a statistical argument, what evidence is useful, how to summarize a distribution, and how to make comparisons between distributions—ALL ACTIVITIES THIS DAY ARE DESIGNED TO FOSTER NORMS OF PARTICIPATION).

H-14-15 may be used as needed to reinforce the construction of and interpretation of boxplots and stemplots. Do not spend much time with these unless it is necessary. If you do use them, H-14 question (g) is most important for interpretation. H-15 (b) introduces the term significant and provides a prelude to the difference between using mean or median as a measure of center.

Activity 1.3 Matching Plots to Variables (35 minutes)

Use H-17 as a warm-up with the group and then assign question 2—Main activity. T-3 is provided for sharing solutions. Listen carefully for groups' reasoning as they assign distributions to contexts. Also, draw attention to the histogram and boxplot concurrent representations. Use the Wrap-Up and Extension questions on H-19 to summarize the activity. Skewness should be part of the discussion as well as mean vs. median, which is greater. . . . May use a *Fathom2* demo at this point—build a distribution that is symmetric and then drag some data values around to see what happens to the mean and median.

Use H-21 (T-4) as a boxplot sense-making exercise. Collect responses and facilitate consensus building if necessary.

Activity 1.4 Standard Deviation and Its Interpretation (1 hour 45 minutes)

(H-22-36) Launch the activity by mentioning that we have been informally looking at distributions and summary statistics. We'll now look at the ever-so-tricky standard deviation. Use H-22 (T-5) to begin the discussion. Ask participants to read H-24 and do #1, 3, 4, 5, 8, 9, 11, 12. Demonstrate the mean absolute deviation (maybe use the data from question 5 to calculate the mean absolute deviation from the median) and use the simulation to demonstrate the influence of outliers on standard deviation in *Fathom2*. This will also introduce teachers to the construction of formulas in *Fathom2*. (mean absolute deviation)

Use H-30 (T-6) as an assessment of ideas about comparing with standard deviations.

Use H-31 (T-7) to discuss the difference between sigma and s. H-32 is a useful guide for helping to select statistics. Briefly discuss the characteristics of good estimators. May use the *Fathom2* demo to demonstrate the fact that mean is more efficient estimator than is median in a unimodal symmetric distribution.

Matching Plots to Statistics (H-33-36; T-8-9) will provide a handy mechanism to check for understanding of standard deviation relative to distributions.

Activity 1.5 Orbital Express Parts II & III (1 hour 5 minutes)

This introduces the randomization test procedure to participants through the physical shuffling of cards to simulate the randomization process. Use H-37-38 to conduct the activity. Select 1 group from which to use the orbital express data. All participants should make cards with the same data so that the class can produce enough resamples to be useful. Collect participants' statistics and summarize in *Fathom*. Facilitate the conversation about what is going on here. Focus on the null hypothesis and the interpretation of the p-value. Encourage participants to voice their ideas and understanding of the process.

Use T-10 to guide participants through the installation of *CPMP-Tools*. Briefly demonstrate a few of the capabilities of the program and then let teachers complete H-39-40.

Closure/Reflection (10 minutes)

Use the written Reflection #1 handout provided for Day 1 to gather participant reactions to initial activities. Assign the readings for the night—especially important is the *Fathom2* reading.

ACTIVITY 2: Technology and the Randomization Test

Overview:

Use of technology is essential to support sound statistical thinking and reasoning. A focus of this session is the introduction to *Fathom2* as a statistical tool. Beginning facility with *Fathom2* will be accomplished through the completion of carefully selected tours. Participants will then use *Fathom2* to model the randomization procedure to continue the Orbital Express activity. Additionally, concepts of randomness, bias, normal distribution, and experimental design will be investigated through their embodiment in several activities.

Goals/Objectives:

Develop facility with *Fathom2*.

Understand that randomness and bias are important statistical ideas which can quite easily be misunderstood.

Introduce or reinforce attributes and characteristics of the normal distribution.

Understand the design of experiments that will allow causation to be established.

Reinforce the randomization test procedure and its interpretation.

Materials to Conduct the Activity:

Fathom2

Graphing Calculators

Random Rectangle Handouts

String and paper bags for Stringing Students Along Activity

Transparencies:	T-1	Random Rectangles
	T-2	Random Number Table
	T-3	Dartboard—Bias & Variability
Handouts:	H-1-15	<i>Fathom2</i> Tours 1, 2, 3, & 6 (from <i>Fathom2</i> Learning Guide—Key Curriculum Press (2005))
	H-16-19	Random Walk
	H-20	General Notes for Permutation/Randomization Tests
	H-21-23	Orbital Express Part IV
	H-24-30	Random Rectangles
	H-31-33	Stringing Students Along
	H-34-37	Streaky Behavior
	H-38	Relationship between bias and variability
	H-39-66	CPMP Course 3 Unit 1: Reasoning and Proof
	R-1***	Franklin, C. & Garfield, J. (2006). <i>The GAISE Project: Developing Statistics Education Guidelines for Grades</i>
Readings:		

- Pre-K-12 and College Courses*. National Council of Teachers of Mathematics Sixty-Eighth Yearbook, Chapter 24.
- R-2*** Shaughnessy, J. M. (2006). *Research on Students' Understanding of Some Big Concepts in Statistics..* National Council of Teachers of Mathematics Sixty-Eighth Yearbook, Chapter 7.
- R-3 **OPTIONAL for A.P. Stats Teachers:**
Bock, D. & Velleman, P. (2006). *Why Variances Add—and Why It Matters*. National Council of Teachers of Mathematics Sixty-Eighth Yearbook, Chapter 25.
- R-4 **OPTIONAL for A.P. Stats Teachers:**
Koehler, M. (2006). *Using Graphing Calculator Simulations in Teaching Statistics*. National Council of Teachers of Mathematics Sixty-Eighth Yearbook, Chapter 18.
- R-5 Orbital Express from Teaching Mathematics with Fathom, Key Curriculum Press (2005)

- *** Indicates a reading assignment required for the next day
- ** Indicates a desired reading assignment for the next day, but if time is short, may read later
- * Great if you can read it before the next session, but not essential
Other readings are for your reference and you may read at your leisure :)

Other Materials

Suggested Teaching Procedures:

Activity 2.1 *CPMP-Tools*—Using Randomization Test and Estimating Mean and Standard Deviation Features (30 Minutes)

During this session, participants will have a further opportunity to practice with *CPMP-Tools*. Encourage participants to play. They can create their own data sets, or perhaps use some from Day 5, H-14-15. Page H-25 from Day 2 contains distributions in which participants may estimate center and standard deviation. If they use *CPMP-Tools* for Course 1, they can investigate the use of the estimate center and spread with the balancing histogram and standard deviation estimator.

May ask: Has your understanding of standard deviation changed over the last two days? If so, how.

Activity 2.2: Tours of *Fathom2* (1 hour, 35 minutes)

Participants should complete Tours 1, 2, and 6. **Working in pairs is suggested throughout the technology portion in order that partners can help each other.** Tour

3 is included as optional for those who might finish early. Completion of these tours should provide participants with enough background to begin to navigate in *Fathom2*.

Because time is so short, have participants complete tours 1 and 2 and then debrief briefly. Tour 6 may be done as a whole group activity if need be. These tours are not meant to produce proficiency with *Fathom2*, rather to introduce capabilities of the software. Proficiency will develop throughout the week.

Activity 2.3: Random Walk & Orbital Express Part IV (45 minutes)

This activity will continue to build proficiency with *Fathom*. First, ask volunteers to physically model a random walk using a coin flip to generate direction. Once the idea is clear, ask students to complete the Random Walk activity (H-16-19). This will provide them an opportunity to create a new case table, establish new cases with the use of a formula, rerandomize the data collection, practice graphing, collecting measures, and interpreting experimental data in a probabilistic setting. If time permits, participants should complete through “Explore More” problem 2. Share and summarize with the group to check for understanding.

Now that participants are relative experts using *Fathom2*, they will use *Fathom2* to conduct a randomization test of the orbital data from the previous day. H-21-23 provides a good initial lesson for the randomization process, scripted but still functional. This should help to further develop proficiency with the software.

Demonstrate how to use the “Stack Attributes,” “Scramble Attribute Values,” and “Collect Measures” features of *Fathom2* so that participants may construct a randomization test template. If all goes well, they should be able to confirm their results from *CPMP-Tools* on Day 1 with the randomization test in *Fathom2*. This activity will begin to acclimate teachers to the resampling hierarchy in *Fathom2*. Reading R-5 is a *Fathom2* version of the Orbital Express Activity, just in case some participants may like to use it in their classrooms.

Activity 2.4: Random Rectangles & Stringing Students Along (45 minutes)

Use the Gorilla video clip to establish the idea that we only see what we’re looking for. . . . This should be quite a conversation starter. Three activities have been included from Activity-Based Statistics. Most likely you will only be able to do the first one. You will likely not have time to complete all three. Random Rectangles and Stringing Students Along are recommended because they highlight the need for random sampling and the potential bias that might result even when a process appears random. Streaky Behavior will likely not be conducted but is included just in case. It would also be a great activity for students. Because of the importance of a thorough understanding of the role of randomness in all of statistics, both of these activities make strong contributions to understanding. You may need to abbreviate some of the activities, however, focus on the main ideas. If necessary, conduct Random Rectangles and then describe the contribution of String Students Along.

Pass out the sheet of rectangles to each participant (face down) (T-1). Then ask them to flip over the page and 1) estimate the average area of the rectangles on the page, and 2) select 5 rectangles they think are representative of the group and then calculate their average area. Collect both sets of information for the group in Fathom. Demonstrate the use of the random number table (H-27; T-2) and the random number generator on the TI-84 and ask participants to randomly select a set of 5 rectangles and calculate the average area. Collect data and compare distributions of the data as well as the mean of the distributions with the actual mean of the area of the rectangles. The issue of sampling bias is the key. Wrap-up question 1 and Extension question 1 are good for assessment (H-28). Use H-38 (T-3) as a quick geometric representation of the relationship between bias and variability. Remind or solicit from participants that good statistics are unbiased and have low variability (efficient).

Activity 2.5: CPMP Course 3 Unit 1: Reasoning and Proof--Design of Experiments and By Chance or From Cause? (1 hour 25 minutes)

This sequence of two investigations begins by introducing concepts of design of experiments that will allow one to confidently determine whether attribution may be determined. The second investigation will reinforce the randomization test procedure that has begun to be developed.

Participants should complete Investigation 1 (will need pennies) and Investigation 2. Investigation 2 can be abbreviated at the front end because of the work we have done previously introducing the randomization test. This lesson will support the interpretation of results from well-designed experiments. Suggested problems are: #3, 4, 5 (only if penny-stacking experiment is done), 6. The summarize the mathematics section (H-52) can provide a good time for a discussion as well as formative assessment of the group's understanding. ("Statistically significant" is used here). Check for Understanding H-53 is also interesting and useful if time permits. Have participants review the remainder of the pages to gain a flavor for the nature of the CPMP materials.

Closure/Reflection (5 minutes)

Assign readings R-1 and R-2 for tomorrow. Use written Reflection #2 for feedback.

ACTIVITY 3: By Chance or By Cause, Continued

Overview:

Participants have been working with comparing distributions using the randomization test. This session will provide the opportunity for them to develop further skill with *Fathom2* as well as new and unique uses of the randomization test for which standard parametric procedures fail. In addition, characteristics of the normal distribution will be investigated.

Goals/Objectives:

Develop flexibility in the use of *Fathom2* to model two-sample comparison problems as well as those that differ from the standard two-sample problem.

Understand the hierarchical nature of generating sampling distributions in *Fathom2*.

Discover relationships between size of a sample and measures of its associated sampling distribution.

Recognize the prevalence of normal distributions in real world data sets and explore applications related to normal distributions.

Materials to Conduct the Activity:

Transparencies:	T-1	Seattle Real Estate Task
	T-2-5	More Randomization Test Problems
	T-6	Certificate
	T-7	Data Table for Certificate & Hand Spans
	T-8	Normal Distribution
	T-9	Sampling Distribution Table
Handouts:	H-1	Seattle Real Estate Task
	H-2-5	Working with <i>Fathom2</i> and the Randomization Test
	H-6-28	CPMP Course 3 Unit 3: Samples and Variation
	H-29	Equation for the normal distribution curve (from CPMP Samples and Variation, Unit 4, p. 18)
	H-30	Mathematics Note: Normal Distribution (from SIMMS Level 4, <i>Nearly Normal</i> , p. 177)
	H-31	Mathematics Note: Binomial Experiment (from SIMMS Level 4, <i>Nearly Normal</i> , p. 166)
	H-32-33	Is There a Relationship Between the Sample Size and the Sampling Distribution Standard Deviation?
Readings:	R-1	Grouws, D. (2003). <i>The Teacher's Role in Teaching</i>

- Mathematics through Problem Solving*. In H. L. Schoen (Ed.), Teaching mathematics through problem solving: Grades 6-12 (pp. 129-142). Reston, VA: NCTM.
- R-2 Rubin, A. & Hammerman, J. (2006). *Understanding Data through New Software Representations*. National Council of Teachers of Mathematics Sixty-Eighth Yearbook, Chapter 17.
- R-3 Groth, R. (2006). *Engaging Students in Authentic Data Analysis*. National Council of Teachers of Mathematics Sixty-Eighth Yearbook, Chapter 4.

Other Materials

Suggested Teaching Procedures:

Activity 3.1: Sharing *Fathom2* Experiences and Trouble-shooting Technology Issues (30 minutes)

Take the first 30 minutes or so and discuss the use of the technology, issues and difficulties. Demonstrate features of *Fathom2* about which participants have questions. This may be a good opportunity to demonstrate the Hospital Problem simulation as participants will be creating sampling distributions later. Also may refer to the R-3 from Day 1 regarding the randomization process.

Activity 3.2: Seattle Real Estate Task (1 hour)

At this time, participants will use the randomization test to compare real estate selling prices for homes in the Seattle area (H-1, T-1). The distributions are skewed, so the issue of an appropriate measure of center is apparent. Computing standard deviation and mean absolute deviation for these distributions will be helpful. Debrief relative to participants' continuing development of ideas about comparing distributions.

Activity 3.3: Other Randomization Test Applications (35 minutes)

Ⓢ implies activity was omitted from the implemented professional development program.

A selection of problems provided for which the randomization test procedure is appropriate (H-2). The first is a matched pairs problem—need to shuffle values by case and then compute statistics. This is easy conceptually, but tricky in Fathom—may be a challenge for someone to implement. Number 2 is a situation when the statistic involves the comparison of ratios for nonparametric tests. Numbers 3a and 3b will challenge participants to analyze the same data in two different ways. Number 4 is an application using the correlation coefficient.

Depending on time, this activity may be jigsawed and processes shared with the group. (#2, 3, 4, and 5 should be accessible in *Fathom2*; #3 is doable in *CPMP-Tools* or *Fathom*; the rest require *Fathom*'s flexibility)

Activity 3.4: Is There a Relationship Between the Sample Size and the Sampling Distribution's Standard Deviation? (1 hour, 45 minutes)

It is expected that this will be a challenging activity (H-32-33, T-9). Groups of 3 or four are suggested for this activity in order that issues of mathematical content knowledge as well as technology may be broadly supported. Participants may need help establishing a beginning population from which to sample, so demonstrate some possibilities. Once the sampling distribution mechanism is set up in *Fathom2*, generating the data will not be difficult, but the importance of this activity is looking for patterns as the sample size changes. Encourage participants to be noticers, especially looking for relationships between the sample and its connection to the sampling distribution. This activity essentially establishes the Central Limit Theorem through mathematical modeling. After the data has been collected, stop to share and summarize across groups. Demonstrate how sliders in *Fathom2* can be used to investigate functions and ask participants to try to find a mathematical model for the relationship between the standard deviation of the sampling distribution and the sample size. A power model is appropriate here, but teachers' experience with modeling may be an issue.

A reflection at the completion of this activity is important for helping teachers to make the connection suggested by the Central Limit Theorem. What have they noticed across the sessions that seems to happen when sampling distributions of means are constructed? (normal-looking distributions. . .)

If time permits, use the *Fathom2* demo of the relationship between standard deviation and range of samples to illustrate that as the sample size increases, the relationship between range and standard deviation from samples becomes less predictable (e.g. the correlation coefficient, though still positive, decreases with sample size increase).

Activity 3.5: CPMP Course 3 Unit 4 Samples and Variation (1 hour, 5 minutes)

The previous activity serves as the launch for the current activity (H-7-31, T-6-8). We will be looking at characteristics of the normal distribution. Begin by using the Think About This Situation on H-7. Pass out copies of the certificate for teachers to use (T-6). Collect the data for the class in *Fathom2* and discuss shape, center, and spread of the distribution (T-7). Have participants in pairs complete Investigation 1, #1-6. Participants will need 3 copies of the normal distribution on H-10 for #3. Use H-30 and T-8 to reinforce the geometric relationships between normal distributions and standard deviations. H-31 is provided as a connection between binomial situation and normal distributions, however, it is only as a resource here.

After the Summarize the Mathematics on H-12, use *Fathom2* to introduce the equation for the normal distribution density curve and the use of slider to investigate its behavior.

This will serve as a launch into Standardizing Scores: Investigation 2 (H-13-16).
Complete as much of Investigation 2 as time permits.

Closure/Reflection (5 minutes)

Reflect on the day's activities. From the randomization test to sampling distributions and the Central Limit Theorem, to the normal curve, we've covered a lot of statistical ground. How are we doing? Assign R-1, 2, 3 for tomorrow. May jigsaw readings & facilitate a short debriefing on Day 4 (maybe share 3 big ideas from each reading).

ACTIVITY 4: Putting It All Together

Overview:

This is the final session in which participants will be asked to reflect upon what they have learned this week and apply some of their ideas in order to assess students' work. They will have an opportunity to continue lingering investigations in *Fathom2* or alternately consider additional statistical investigations for which they may design ways in which *Fathom2* may be incorporated into the investigations.

Goals/Objectives:

Apply statistical knowledge to assess student work in an authentic data analysis setting.

Understand the application of the randomization test to an authentic statistical investigation

Connect the relationship between what has been learned about the relative predictability of sampling from a simulation perspective to the statistical work of working backwards from a sample to a population

Materials to Conduct the Activity:

Transparencies:	T-1-3	Matching Samples to Density Curves
Handouts:	H-1-3	Physicians' Reactions to Patient Size
	H-4-19	Assessing Students' Responses to Authentic Data Analysis Problems
	H-20	The GAISE Framework
	H-21-22	Matching Samples to Density Curves
Readings:	R-1	Copes, L. & Shager, N. K. (2003). <i>Phasing Problem-Based Teaching into a Traditional Education Environment</i> . In H. L. Schoen (Ed.), <i>Teaching mathematics through problem solving: Grades 6-12</i> (pp. 195-206). Reston, VA: NCTM.
	R-2	Stein, M. K. & Boaler, J. (2003). <i>Teaching Mathematics through Problem Solving: Research Perspectives</i> . In H. L. Schoen (Ed.), <i>Teaching mathematics through problem solving: Grades 6-12</i> (pp. 245-256). Reston, VA: NCTM.
	R-3	Rasmussen, C., Yackel, E., King, K. (2003). <i>Social and Sociomathematical Norms in the Mathematics Classroom</i> . In H. L. Schoen (Ed.), <i>Teaching mathematics through problem solving: Grades 6-12</i> (pp. 143-154). Reston, VA: NCTM.
	R-4	Konold, C., & Higgins, T. L. (2003). <i>Reasoning about data</i> . In J. Kilpatrick, W. G. Martin & D. Schifter (Eds.),

R-5	SIMMS Integrated Mathematics Level 2, 2 nd Edition (2003)
R-6	SIMMS Integrated Mathematics Level 4, 2 nd Edition (2003), Nearly Normal
R-7	SIMMS Integrated Mathematics Level 4, 2 nd Edition (2003), Confidence Builder

Activity 4.4: Matching Samples to Density Curves (20 minutes)

Use H-21-23 and T-1-3 to reinforce the idea that larger samples provide more information about the population. This activity should really make that clear. Shapes of some density curves should be mentioned—normal, uniform, skewed, bimodal. Ask participants to reflect on the differences between samples and sampling distributions of some statistic.

Activity 4.5: Making Connections—From Sample to Population (30 minutes)

Most of the activities for the week involved beginning with a population, taking multiple samples to create sampling distributions, or alternately using resampling with the randomization test given two distributions. Ordinarily, however, one usually has only 1 sample from which to infer something about the population. From this should come the notion that larger samples provide more information, narrower sampling distributions, and more precision regarding working backwards to the population mean or other desired parameter. R-5,6, & 7 are units from SIMMS materials that have been included as resources that might stimulate further investigation into statistical areas—designing surveys, working with normal distributions, and confidence intervals.

Activity 4.6: Closure/Reflection (x minutes)

Looking back on the week, generate a list of big ideas that have been investigated. Look back at the Standards/GLEs/MCF relative to expectations for kids. Encourage teachers to share their thoughts with the group.

It will be important to walk back through all of the resources that have been provided in the notebook for participants. Draw their attention to supplemental activities and readings that were included for them to investigate on their own. Thank them for their hard work this week and provide encouragement to continue their statistical learning journey!!

Post-Assessment (1 hour, 30 minutes)

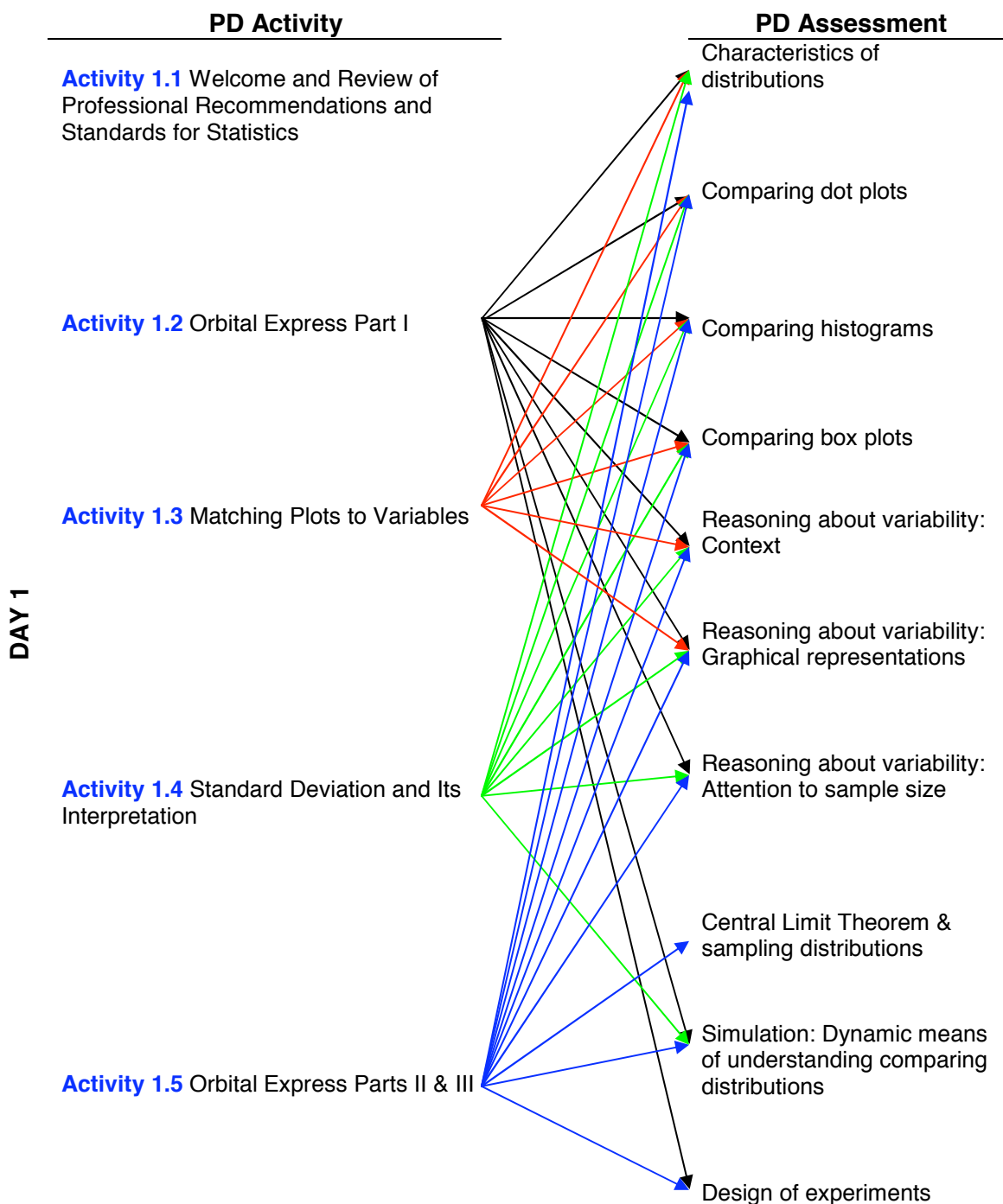
Tickets Out The Door (works well to use the Stipend Reimbursement Form in exchange for necessary documentation):

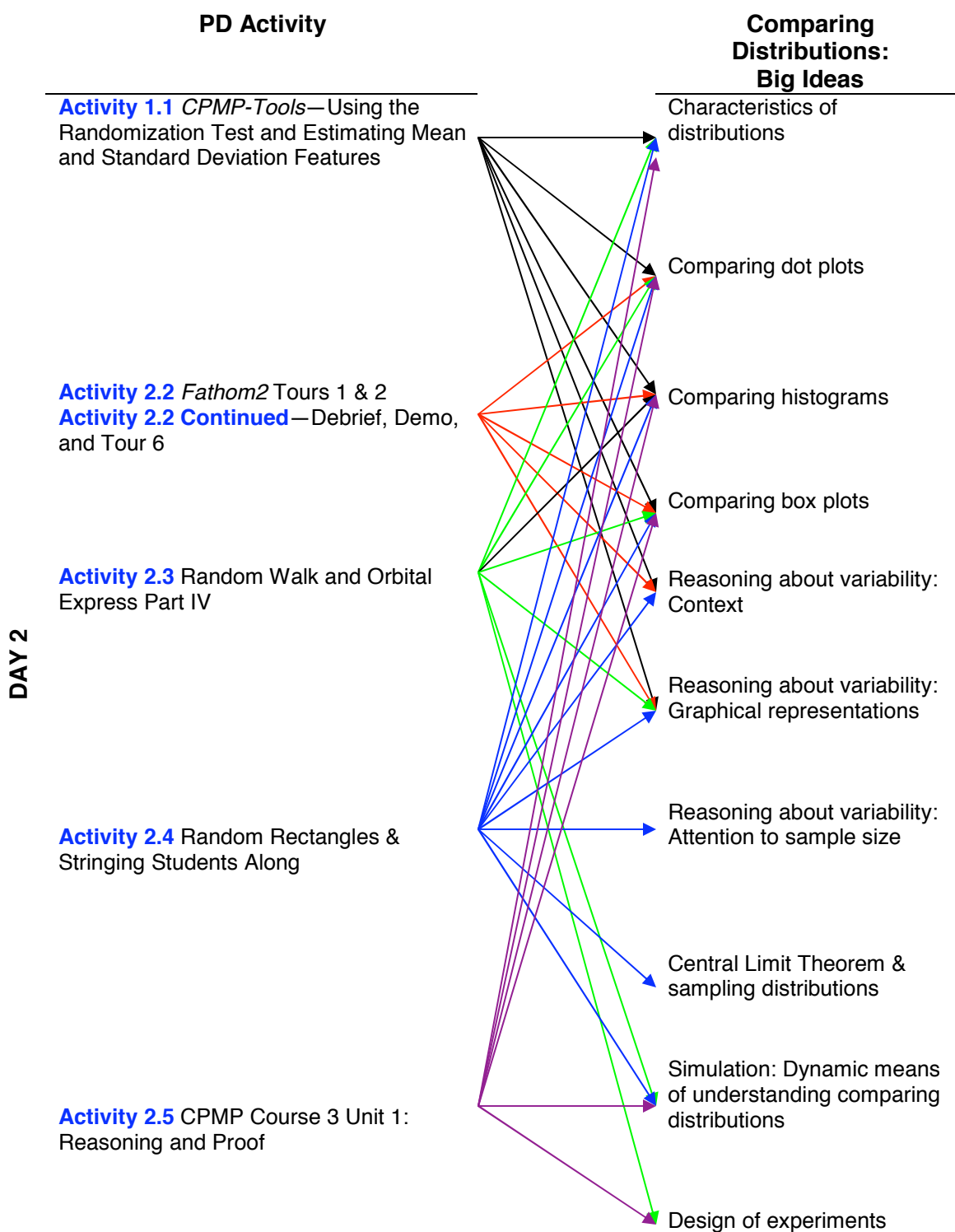
- 1) Reflections
- 2) Evaluation
- 3) Post-test
- 4) Teacher Schedule

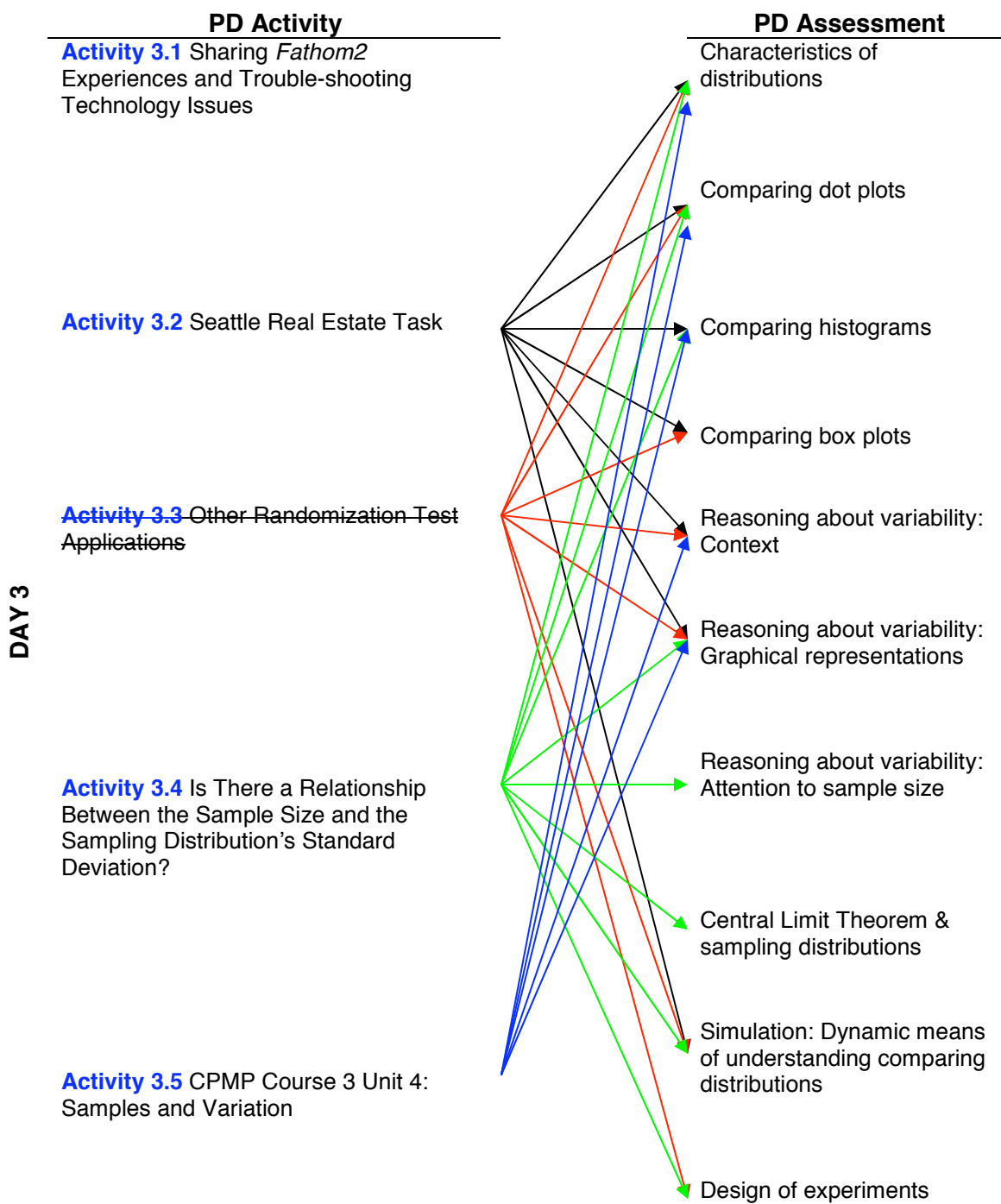
Appendix H

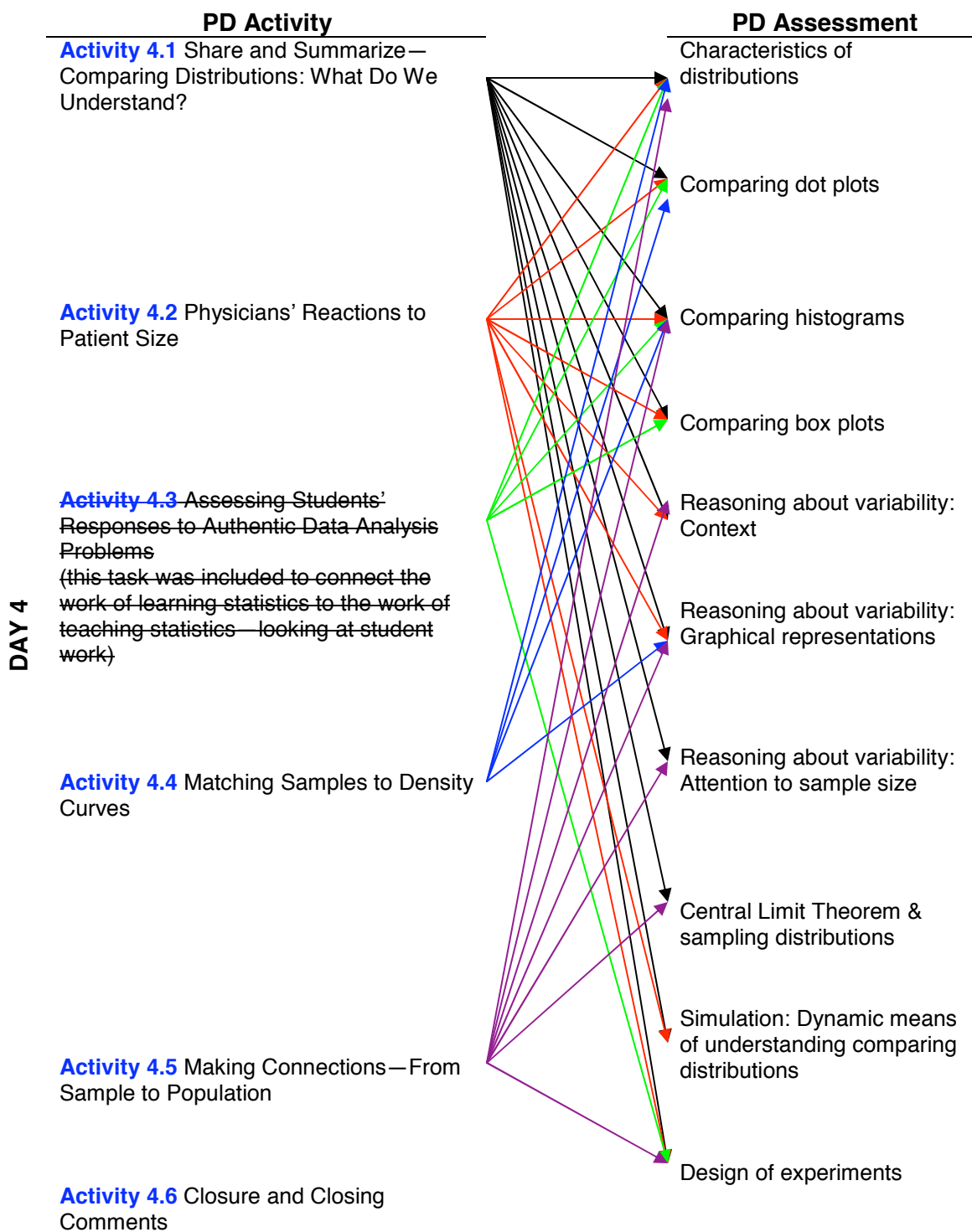
Mapping Professional Development Activities to Assessments

Mapping Professional Development Activities to Big Ideas of Comparing Distributions









Appendix I

Human Subjects Institutional Review Board Approval

WESTERN MICHIGAN UNIVERSITY



Human Subjects Institutional Review Board

Date: June 8, 2006

To: Steven Ziebarth, Principal Investigator
Mark Jenness, Co-Principal Investigator
Sandra Madden, Student Investigator for dissertation

From: Amy Naugle, Ph.D., Chair

Re: HSIRB Project Number: 06-05-22

This letter will serve as confirmation that your research project entitled "High School Mathematics Teachers' Evolving Knowledge of Comparing Distributions" has been **approved** under the **expedited** category of review by the Human Subjects Institutional Review Board. The conditions and duration of this approval are specified in the Policies of Western Michigan University. You may now begin to implement the research as described in the application.

Please note that you may **only** conduct this research exactly in the form it was approved. You must seek specific board approval for any changes in this project. You must also seek reapproval if the project extends beyond the termination date noted below. In addition if there are any unanticipated adverse reactions or unanticipated events associated with the conduct of this research, you should immediately suspend the project and contact the Chair of the HSIRB for consultation.

The Board wishes you success in the pursuit of your research goals.

Approval Termination: June 8, 2007

Walwood Hall, Kalamazoo, MI 49008-5456
PHONE: (269) 387-8293 FAX: (269) 387-8276