The Dissertation Committee for Katie M. Makar
certifies that this is the approved version of the following dissertation:

Developing statistical inquiry:

Prospective secondary mathematics and science teachers'

investigations of equity and fairness

through analysis of accountability data

Committee:

_____
Jere Confrey, Supervisor

_____
Jill A. Marshall, co-Supervisor

_____
Martha K. Smith

_____
Thomas W. Sager

_____
Jennifer C. Smith

Developing statistical inquiry:
Prospective secondary mathematics and science teachers' investigations of equity and fairness through analysis of accountability data

by

Katie M. Makar, B.A., M.A.

Dissertation

Presented to the Faculty of the Graduate School of

the University of Texas at Austin

In Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May, 2004

Dedication

This dissertation is dedicated to four caring, insightful, and progressive women that have been my intellectual mothers over nearly three decades of my life. Without these women, I would not be on this journey. They saw potential beyond the life I imagined for myself and encouraged me seek a path toward its realization.

Sue Ann McGraw
my high school mathematics teacher

Dr. Barbara LiSanti
my college mathematics professor

Dr. Robin Heslip
my school curriculum coordinator and teaching mentor

and

Dr. Jere Confrey
educational visionary and my PhD supervisor

Acknowledgements

I would first like to acknowledge the support I have received through two grants from the National Science Foundation (DUE-9953187 and ESR-9816023) over the past four years that has subsidized my tuition, stipend, health insurance, resources, and professional development. In particular, the funding I received from the Collaborative for Excellence in Teacher Preparation, a program from the National Science Foundation supported me during the data collection and writing stages. Without the support of these grants and the generosity of the Principle Investigators (PIs), Prof. Jere Confrey and Prof. Michael Marder, I would have not been able to concentrate on my degree, and certainly not been able to afford it. I would also like to thank the UTeach program in the College of Natural Sciences and College of Education at the University of Texas at Austin for their willingness and support in allowing me to conduct this study and their flexibility for permitting experimentation with the course in which the study took place.

I am forever indebted to Jere for her constant encouragement, incredibly high expectations, relentless probing, enduring patience with my questions and stubbornness, and unbelievable confidence in me, right from the beginning. Her undying energy, insight, brilliance, scholarship, and zest for the cause of improving education can never be matched, but has been an example I will always aspire to. She told me not how to think, but challenged me to rethink. These four years have been the hardest and most rewarding years of my life; under Jere I've had to question my assumptions about the world and reconstruct my beliefs about reality and truth and perspective. I'll never forget arguing about whether a chair is real. I feel that I'm coming away a different person intellectually and emotionally. I'm just beginning to learn to fight for my own ideas, challenge those of others, and seek out honest critique because of Jere's mentorship. When she scrawled "Go, Girl!" on my notepad as I challenged senior researchers with shaky voice at a Harvard conference, I felt such a boost of confidence. I've fallen in love with research, philosophy, epistemology, and equity – areas I had never explored nor valued before I met Jere. It is with tremendous respect and humility that I am now able to call her a colleague.

I could have never gotten through this last year without Dr. Jill Marshall. Jill read, re-read and read again every draft I wrote and responded within hours with feedback, constructive criticism, and support. Her final edit of the entire document was

a tremendous gift. If she didn't hear from me for a few days she was quick to email and ask me how things were going with offers to meet and talk through frustrations or read half-written ideas. She was always willing to talk through ideas even before she was on faculty and I learned a lot being a co-author with her. I always felt like I could be completely honest with her about my shortcomings without any fear of judgment. She has earned my respect and endearment like an older sister. She even appreciates Dylan.

I wish to thank the other three members of my committee, Prof. Tom Sager, Dr. Jennifer Smith, and especially Prof. Martha Smith. Martha Smith gave me wonderful feedback on my early drafts and pushed me to rethink some of my notions about variation and informal terminology. She was never afraid to critique my ideas, always with my best interest at heart, and gently rebuke me to maintain high levels of scholarship when I was overly zealous in my claims or overly critical of ideas that differed from my own. She is my ideal of a statistician.

I am so grateful to the preservice teachers who were willing to tolerate a pretest on statistics the very first day of class, patiently endure being videotaped and interviewed and probed. I learned so much from them and will always appreciate their honesty and willingness to help me learn about teaching teachers. I also am very appreciative of the practicing teachers that worked with me on the pilot project, especially Michael and Michelle, who provided me with so much to think and write about, even opening their classrooms to me a year later for weeks of observations and more interviews. They let me cut my teeth as a researcher on them and I will not forget their kindness and generosity with their time.

My graduate school friends Jennifer, David, Sibel, and Erica created a wonderful support system for me, without which I doubt I could have survived. Jennifer was so quick to help me learn the ropes and offer advice on classes and graduate school survival, impossibly balanced with motherhood. Insightful historian and philosopher David, who shared my new love for epistemology (except that he understood what he was talking about), spent hours in the office with me debating ideas, venting frustrations, and sharing inspiring articles; he has been a wonderful friend these four years. Sibel, my co-statistician, has been such a godsend to be able to ponder emerging and half-formed statistical ideas with, even from St. Louis. My daughter will always think of her as Sibel-massie (Aunt Sibel). Erica, especially this past year, was such a

good friend who even in a very difficult stage of her own life took time out to cry with me and share frustrations over lunches in Georgetown. Those times are cherished.

I also want to thank my friend Melissa Tothero who always had a way of rephrasing things so that what felt like a tremendous obstacle seemed smaller, and little successes seemed like momentous occasions. And my dear friend Lisa Kridner, the best first grade teacher I've ever known, who reminded me how much I love to teach and to live with the constant struggle of never being completely satisfied with it.

My mentors and international colleagues in the statistics education research community, especially the SRTL'ers, have been wonderfully supportive and encouraging in my research pursuits these past few years. I look forward to many more years of sharing and building ideas together. These have got to be the nicest group of researchers that ever existed. I especially want to thank Joan Garfield, Dani Ben-Zvi, Michael Shaughnessy, Dan Canada, Jane Watson, Chris Reading, Beth Chance, Rolf Biehler, Graham Jones, Iddo Gal, Cliff Konold, Andee Rubin, and especially Arthur Bakker, for their advice and stimulating statistical discussions in person and over email.

I want to also lovingly thank my parents, my dear mother Marilyn Roberts and dear father the late James Roberts. They have had to endure my world travels and independent perspective, but have continued to encourage me to do what I love. My mother's excitement about seeing me graduate with my Ph.D. helped me get through these last few weeks.

Most importantly, I wish to express my utmost indebtedness to my wonderful husband Sanjay and daughter Keya who allowed me to drag them 10,000 miles away from our idyllic tropical lifestyle so I could pursue a dream and a new career. My beloved husband endured extra long hours working to pay the bills and put up with my countless conference trips, disappearances to go study, neglect, papers and books forever all over the house, and unpredictable grumpy moods without ever questioning why we were here. He teaches me time and again about patience and tolerance and reminds me about what is really important. My dear Keya has been so patient in ways that no little girl should have to be when mommy was too tired or too busy to play or talk. Her cheerfulness and excitement about learning gave me such energy and pleasure. I am forever grateful to you two and so excited as we move into this new phase of life together in Australia.

Developing statistical inquiry:

Prospective secondary mathematics and science teachers' investigations of equity and fairness through analysis of accountability data

Publication No. _____

Katie M. Makar, Ph.D.
The University of Texas at Austin, 2004

Supervisor: Jere Confrey
Co-supervisor: Jill Marshall

Concerns about equity in the ways that schools are using the data from the results of their students' state-mandated exams (Confrey & Makar, in press) prompted this mixed-method study, based on the model of Design Research (Cobb et al., 2003). The study was conducted to provide insight into the ways that understanding of the statistical concepts of variation and distribution, developed in the context of learning about equity and assessment, can allow prospective teachers to broaden their understanding of equity and gain experience with conducting an inquiry of an ill-structured problem through the use of data generated by high-stakes tests to investigate equity and fairness in the accountability system. The study took place in an innovative one-semester course for preservice teachers designed to support and develop understanding of equity and fairness in accountability through data-based statistical inquiry (Confrey, Makar, and Kazak, 2004). The prospective teachers' investigations were conducted using Fathom Dynamic Statistics (Finzer, 2001), a learning software built for novice data analysts which emphasizes visualization and building inferential

thinking through highlighting relationships between multiple variable displays. Semi-structured investigations during the course led up to a three-week self-designed inquiry project in which the prospective teachers used data to investigate an area of interest to them about equity in accountability, communicating their findings both orally and as a written paper.

Results from the study provide insight into prospective teachers' experiences of conducting inquiry of ill-structured problems and their struggle with articulating beliefs of equity. The study also reports how statistical concepts documented in structured settings showed that the subjects developed rich conceptions of variation and distribution, but that the application of these concepts as evidence in their inquiry of an ill-structured problem was more challenging for them. No correlation was found between the level of statistical evidence in the structured and open-ended inquiry settings, however there was a significant correlation between prospective teachers degree of engagement with their topic of inquiry and the depth of statistical evidence they used, particularly for minority students. Implications and suggestions for improving the preparation of teachers in the areas of statistical reasoning, inquiry, equity, and interpreting assessment data are provided.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

This is a critical time in mathematics education. Teachers are struggling to change the nature of their instruction to respond to the visions put forth by the National Council of Teachers of Mathematics (NCTM, 1989; 2000) and articulated in the National Science Education Standards (NRC, 1996). At the same time, schools are under pressure by politicians and the public to improve student performance on state-mandated assessments, which are generally not well aligned with NCTM's vision. Politicians argue that the data on student performance can be used to improve instruction by providing feedback on content strands where students are systematically weak while simultaneously attending to gaps in performance by students who systematically under-perform on standardized tests. The research documented in this study will be useful primarily to assist researchers and teacher educators seeking to improve instructional practice through the use of data from standardized testing.

This study was designed to provide insight into the ways that a deep, conceptual understanding of the statistical concepts of variation and distribution can allow teachers to use the data generated by the high-stakes tests to seek understanding of equity and fairness in the accountability system. The study described in this dissertation takes place in an innovative one-semester course for preservice teachers designed to support and develop understanding of equity and fairness in accountability through data-based statistical inquiry. The preservice teachers conducted investigations using Fathom Dynamic Statistics™ (Finzer, 2001), a learning software built for novice data analysts which emphasizes visualization and building inferential thinking through highlighting relationships between multiple variable displays. Semi-structured investigations during the course led up to a three-week inquiry project in which they developed, investigated, and presented their findings. The results of this study can provide researchers and teacher educators insight into the interactions between teachers' conceptual understanding of variation and distribution and their engagement with statistical inquiry into equity and fairness in accountability. Furthermore, the description of the innovative assessment course given here and elsewhere (Confrey, Makar, & Kazak, 2004) can assist preservice teacher educators in understanding the kinds of results, insights, and difficulties that preservice teachers experience when they are provided the opportunity to simultaneously develop their understanding of the purpose and limitations of

standardized testing, engagement with inquiry, proclivity towards equity, statistical content knowledge, and facility with dynamic technology.

1.1 CONCEPTUALIZING THE PROBLEM

Teachers are increasingly under scrutiny to respond to two competing reform movements, both of which articulate higher "standards" and argue that their approach attends to equity. Confrey (in preparation; Confrey, Bell, & Carrejo, 2001) calls the unfortunate location of schools and students under this set of conflicting reforms *systemic crossfire*. On one side, the mathematics education reform, as envisioned by the NCTM Standards, is asking that teachers change the curriculum they teach, their method of instruction, and their assessment practices. The changes the Standards are advocating require a substantial change in practice that requires an epistemological shift in teachers' beliefs about teaching and learning of mathematics, a tremendous personal commitment to long-term change, a deep understanding of mathematics, a willingness to take risks, and years of experience to develop. The payoff for their students, they are told, is a deeper understanding of mathematics, better preparation for an increasingly complex and technologically-driven world, and greater engagement, particularly those who have traditionally been disproportionately under-represented in higher levels of mathematics. This approach has been shown to be effective on a small scale, but difficult and expensive to scale up to a larger population of schools.

On the other side, politicians promote their commitment to higher standards of learning and argue that the path to reaching higher standards is by mandating improvement and simultaneously using high-stakes tests to ensure that their mandates are being executed. Confrey (in preparation) calls this the "bookends" model of improvement – increased learning is mandated and outcomes are tested but schools are left to figure out the rest. Because the state-mandated tests are generally skills-based, the reform put forth by the accountability system is easier (and cheaper) to respond to. *No Child Left Behind*, put forth by President Bush (U.S. Department of Education, 2001; 2002), leaves state legislatures with little choice but to pressure schools into implementing the mandates. The consequences of not attending to the accountability system are high—loss of funding and threat of closure or state take-over for schools that do not conform.

Texas, where much of Bush's educational reform was developed, has a high-stakes accountability system. Until 2002, the Texas Assessment of Academic Skills

(TAAS) assessed student performance on the state curriculum. Schools which did not meet the state's criteria for performance on TAAS were given a "low-performing" rating and subject to state sanctions if they did not improve scores within two years. Schools have responded to the accountability system by looking for strategies to improve test scores. In the last five years, a plethora of books and "improvement programs" have emerged from the private sector to assist schools in using the data of their students' performance on state tests to improve their test results, particularly for minority students (for example, Johnson, 2002; Love, 2002, 2003; Scheurich & Skrla, 2003; Schmoker, 1996). Although most of these books focus on positive, long-term strategies for improvement (e.g., creating a school culture of inquiry, attention to over-tracking of minorities, breaking down stereotypes) almost all of these publications advocate that schools disaggregate their data by ethnicity, gender, and economic status, particularly since almost every state reports on and holds schools accountable for their disaggregated test results.

1.2 TWO CRITICAL EXAMPLES

The problem with a focus on disaggregating data is that in the rush to improve their scores, many schools are pressured to focus on short-term gains rather than long-term improvement. This can lead to stereotyping and strategies that, while they increase percentage of students who pass the state test, are questionable ethically. Poignant examples come from two local urban high schools where many of the preservice teachers at our university conduct their practice teaching. One of these cases eventually led to the study conducted in this dissertation.

In both examples, the schools were rated low-performing when fewer than 50% of their small population of African-American students passed the mathematics portion of the state exam. The response by the first school was to mandate *all* of their tenth grade African-American students, regardless of test performance, to regular lunchtime tutoring for the school year (Kurtz, 1999). The result was that the subgroup posted higher scores that year and the school was held up as a pillar in the district for their solution to the "problem" of poor achievement by their African-American students.

The other high school was part of a five-year partnership between its mathematics department and our research group, the Systemic Research Collaborative for Education in Mathematics, Science, and Technology (SYRCE). When the school fell low-performing, the partnership was ended suddenly despite the improvement in

teaching practices and student scores over the time of the partnership. Instead, the school chose to focus on "fixing" the low-performing group with additional test preparation mandated in mathematics classes and the instigation of a pull-out program of its minority students for extra tutoring.

Further analysis of the school's data by SYRCE revealed several questionable decisions by the school administration (Confrey & Makar, in press): (1) they did not consider the effect of the increase in passing standards by the state that year on the long-term improvement trajectory of the students at the school, nor consider that the passing rate of the under-performing subgroup was the second highest in the past five years; (2) they did not take into account the possibility of natural variation along this trajectory—the drop in percentage passing by the small subgroup (n = 31) was within one-standard deviation of their predicted performance based on their performance trajectory over several years; (3) the approach taken by the school was strongly oriented towards "fixing" the students rather than looking for underlying problems at the school that may have resulted in the subgroup's poor performance; (4) the school took a reactive, rather than proactive, stance towards improving their scores, not considering, for example, the larger curricular goals of the school; (5) the administration did not consider the distribution of scores of the subgroup that under-performed; had they done so, they would have seen that if *one* student had answered one or two additional questions correctly, the school would not have been labeled low-performing; and (6) finally, the school took on a number of questionable strategies to avoid low-performance in the following year, such as focusing test preparation on the under-performing subgroup, conducting pull-out programs (particularly for minority students), and by ensuring that the number of African-Americans tested remained below the required number (n = 30), for example, by reassignment of at least one student to special education status without his or his parents' knowledge.

Unfortunately, in both of the cases described above, the schools focused on increasing their percent passing rates, rather than focus on improvement of overall student learning. Their short-term reaction to "using data" turned into a focus on "fixing" the low-performing subgroup rather than probing underlying reasons why these subgroups may have been neglected. Our research team became concerned that in the rush to improve test scores, particularly of minority students, that schools can take on strategies that lead to stereotyping and removal of educational opportunities in

exchange for increased ratings by the school. This led us to "question the implications and approaches drawn from the disaggregation of data and the design of the accountability system in the case of small populations" (p. 9, Confrey & Makar, in press).

The outcome of our own investigations at SYRCE over the next three years included an approach to analysis of school data that focused on equity in ways that were meant to break down stereotypes rather than reinforce them. This approach is based on combining strong contextual knowledge of testing, inquiry-based statistical analysis emphasizing a visual focus on distributions, and the use of dynamic statistical software, and resulted in the setting for the pilot and dissertation study described here. Because the study conducted in this dissertation was situated within SYRCE's larger research agenda, it is important to include further background about the research group to understand better how this study fits within the larger vision at SYRCE.

1.3 THE SYSTEMIC RESEARCH COLLABORATIVE FOR EDUCATION IN MATHEMATICS, SCIENCE, AND TECHNOLOGY (SYRCE)

The Systemic Research Collaborative for Education in Mathematics, Science, and Technology (SYRCE) directed by Professor Jere Confrey (formerly at the University of Texas at Austin), developed a model of systemic reform (Figure 1.1) to guide its professional development work (Confrey, Castro-Filho, & Wilhelm, 2000). In this model, the outcome of Standards-based instruction can be measured by using student data and artifacts, which in turn should drive professional development, which can work to improve teacher knowledge and sense of community, further improving reform-based instruction. The work carried out by the team at SYRCE examined partnerships between universities and schools through a type of research called *implementation research* (Confrey et al., 2000) was unique in that it was:

1. *systemic* in nature by studying classrooms not as independent entities, but situated within a larger system of schooling with the intention of generalizing beyond the setting under study;
2. developed to directly inform *practice* as well as the research community;
3. *embedded* within functioning classrooms so that the research was situated within the context of the workings of schools with the intention of having direct impact on the instructional core of those classrooms; and finally,

4. focused on the *process* of moving through improvement rather than a snapshot in time which does not inform improvement strategies.

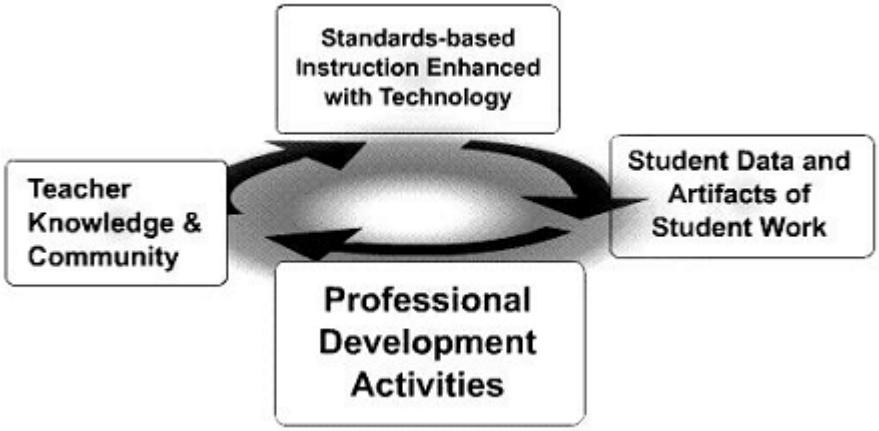


Figure 1.1: The SYRCE model of Systemic Reform used to drive professional development.

The study described in this dissertation came out of other ongoing research projects being conducted by the team at SYRCE over the years 1996 - 2003. Other research conducted by the research group in relation to the SYRCE model has included:

1. Examination of urban high school teachers' professional development which focused on building their content-knowledge of algebra and functions, supported their move to standards-based instruction, and documented improved teacher community (Castro-Filho, 2000; Confrey, in preparation; Confrey et al., 2000; LaChance, 1999; LaChance & Confrey, 2003).
2. Discussion of the tension, or *systemic crossfire*, between the competing trends of reform and its measurement of progress by state mandated multiple-choice tests. This tension is exemplified by documenting how an urban high school, with a diverse student population, in which the research team had a five-year partnership, dismantled their practice towards Standards-based instruction in reaction to being designated low-performing by the state. This was in spite of attempts by the team to point out the progress that had been made during the partnership and that the reason for being labeled low-performing (due to a small subgroup, n = 31, of African-American students falling below 50% passing) was not unexpected within their long-term trajectory of performance and could be

due to chance variation from a small subgroup. In addition, the team documented how, in order to increase of the percentage of students reported passing on TAAS, the school took part in questionable strategies such as focusing on test-preparation and pull-out programs of minority students, evidence suggesting reassignment of a student at risk of failure to special education status without his knowledge, and special attention to improving scores of *bubble kids*—those whose previous scores are close to the passing standard (Confrey, in preparation; Confrey, Bell et al., 2001; Confrey & Makar, in press).

3. An investigation and challenge to the validity of the practice of providing teachers results of their students' raw TAAS scores disaggregated by content strands when psychometric methods of scaling to ensure equivalent difficulty is done on the entire test (Confrey, 2002b; in preparation; Confrey & Carrejo, 2002a; 2002b).
4. A fourth project conducted by the research team examined the aspect of the SYRCE model which focuses on teachers' feedback from student assessment results for professional growth, and served as a pilot study for this dissertation. The study involved a six-month partnership with an urban middle school to help teachers improve their understanding of students' TAAS results (Confrey & Makar, 2002; Confrey, Makar, & Nicholson, 2001; Makar & Confrey, 2002; in press). The results of the pilot study, discussed further in the next chapter, found that through experience in data investigations and conducting their own inquiries into student test data, the middle school teachers in the project not only gained statistical understanding, but were better able to develop more inquiry-based habits of mind. The teachers developed a much richer understanding of questions that could be investigated with data. In the beginning of the study, their data questions focused on simple well-defined questions that would produce yes or no answers, or ones that were too complex to measure with available data. By the end of the pilot study, however, the teachers' inquiry projects displayed a much deeper level of thinking about what is possible to uncover in data. Furthermore, teachers became comfortable using dynamic learning-oriented software, and there was some evidence that through their data-based inquiry experiences, the teachers developed a greater sense of

understanding about the need to nurture all of their students. The opportunity to modify the professional development conducted in the pilot study for use in a university course for prospective mathematics and science teachers seemed an ideal opportunity to carry out this dissertation study, particularly given the calls for increased work in equity, assessment, inquiry, and content knowledge in preservice programs.

1.4 OVERVIEW OF THIS STUDY

The subjects in the study were eighteen undergraduate math and science majors enrolled in a one-semester course in the UTeach program, a joint teacher education program designed for secondary preservice mathematics and science teachers in the College of Education and the College of Natural Sciences at the University of Texas at Austin. The course was modified from its original form for the purpose of this dissertation. The plan of the study followed the model of educational inquiry called *design research* (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003; Confrey, 2002a; Edelson, 2002), where the emphasis is placed on its utility, iterative design process, and innovative curricular treatment. Both quantitative and qualitative data was collected during the course, including pre-post test, interviews, and course artifacts.

The course focused on four major themes: assessment, classroom instruction, equity, and inquiry. These themes were also in the original course, but the revised version used for this study organized these themes differently, interwoven with statistical inquiry of assessment data. The capstone of the course was a three-week inquiry project in which the prospective teachers individually (or in pairs) designed and conducted their own data-based investigation of an issue of equity and fairness in accountability. Within a context that supported its development, it was conjectured that the preservice teachers would use the understanding of variation and distribution that they gained to argue for more equitable treatment of students by schools.

1.5 UNDERLYING ASSUMPTIONS

The dissertation study described in this document was driven by the assumption that if teachers (preservice or practicing) are given the opportunity to gain a strong conceptual foundation of statistical variation and distribution through inquiry and analysis of assessment data, they will develop a greater proclivity towards equity and fairness in accountability and testing. There are a number of themes which underlie this

assumption and form the basis for the revised course conducted for this study. For one, there is an assumption that a deep contextual understanding of assessment and accountability are critical in order to observe the connection between statistical evidence and equity in high-stakes assessment practices. Second, that the preservice teachers need to develop a broader perspective of and lens towards seeing equity. Third, that the required understanding of variation and distribution necessitates a very different approach and conceptual focus than is usually given in a standard statistics course. This conceptual understanding, in fact, is assumed to be critical for the preservice teachers to "see" students that are frequently neglected, or overly treated, by the accountability system. It is further assumed that the facility with technology lent further strength to the teachers' ability to conduct their investigations; in fact, I believe it would have been impossible without it. The interactions between these four themes—contextual knowledge of assessment through data investigations and inquiry, statistical understanding, focus on equity, and technology use—are so intertwined that it is difficult to separate them. Yet, all are necessary to the revised course and the dissertation study.

While this set of assumptions is based on experiences in the research team and with teachers in the pilot study, it is openly acknowledged that this assumption has not been tested with a controlled experiment. Before testing this assumption more rigorously, I determined it was critical to first gain a deeper understanding of the relationship between the prospective teachers' understanding of variation and distribution and data-based inquiry of issues of equity and fairness in accountability, as well as allow the treatment to be adapted for use in a new setting, with secondary preservice teachers. Furthermore, there is an inherent danger, because of the complexity of the conjecture, of trying to study all of these areas in depth and by doing so losing focus. Any one of these areas, and any interactions between them, are important areas deserving their own research, as will be argued in the literature review.

1.6 FOCUS OF THE STUDY

The focus of this study will be on the interaction, within the context of the assumptions underlying the course, between the preservice teachers' engagement with issues of equity and accountability in their inquiry project and their use of variation and distribution as evidence in their inquiry into this topic. With that focus, my main research question is:

> In a preservice course created to support learning about assessment, technology-driven data analysis, equity, and inquiry, how do prospective teachers use the concepts of variation and distribution to support their understanding of issues of equity and fairness in testing?

This question assumes that they *do* use variation and distribution to support this understanding of equity, and one of the elements under study in *how* they use statistics is whether there is evidence of affordances and constraints within the task, their motivation, and the classroom environment that support their use of statistics as evidence in their inquiry. As discussed above, many factors could be examined in depth in order to narrow the scope of the main research question. I saw four major elements that would influence the preservice teachers' likelihood of using statistical evidence in their data inquiries into fairness in testing: (1) their understanding of statistics, particularly variation and distribution, including ways that they articulated this understanding; (2) their use of the technology to analyze the data; (3) their beliefs about equity; and (4) the evidence they sought to support their inquiry. To help narrow the focus to consider these factors, four sub-questions were developed:

1. What level and types of understanding of the concepts of distribution and variation were learned? How did the teachers express this understanding in practice?
2. How was the technology used in relation to the students' inquiries? What behaviors did the prospective teachers exhibit in using the technology in a semi-structured investigation?
3. What can be said about preservice teachers' understanding of equity from their structured and ill-structured inquiry activities?
4. What is the interplay between the preservice teachers' statistical reasoning and the depth and breadth of self-designed inquiry into complex, ill-structured problems?

These four questions will be the major focus of Chapters 4 and 5, which document the results of analysis of the statistics pre-post test, interviews involving structured investigations, reflection papers on issues of equity written during the course, and their inquiry projects.

1.7 POTENTIAL BENEFITS

Several benefits to the research community and practioners could possibly come out of this study. In the area of research, there is the potential for additional insight into the statistical reasoning the teachers used in a complex and compelling context. Additionally, those interested in research on equity may gain insight into teachers' understanding of equity and its potential enhancement with the tools and experience of statistical inquiry of equity. Those interested in studying teachers' use of inquiry may find this an interesting study to examine the context and product of the prospective teachers' inquiries. For practitioners interested in improving teacher education, this study provides insight into a form of professional development that could simultaneously enhance preservice teachers' content knowledge and identity as professionals, facility with a dynamic technology package that can be used with students, and also allow preservice teachers to experience being learners in a reform-based, inquiry-rich, technology-enhanced environment with authentic content. Additionally, for preservice teacher education programs, the unique combination of the themes of the course described here begin to paint potential for a more connected, authentic preservice curriculum. Professionals working with practicing teachers can also see potential with that population to see the kinds of results and understandings that teachers can uncover if given the time, opportunity, and tools needed to go beyond superficial analysis on their students' testing results, although this study focuses on preservice teachers. The examples of investigations and inquiry topics the preservice teachers chose described in this study may provide insight for policymakers to see a need to focus more attention on distributions of student performance, perhaps by requiring that schools attend to improvement for each quartile of student results.

1.8 OUTLINE OF THE DISSERTATION CHAPTERS

The interactions that this research documents between preservice teachers' understanding of statistical concepts, reflection into issues of equity in testing, and experience in conducting their own inquiry provides evidence of ways in which these skills and understandings can play out *in practice* in a preservice teacher course. This introductory chapter set the stage for why this study is important and timely in the current craze of data use by schools, as well as how this study is situated within a larger research agenda of improved professional development in systemic reform. Chapter 2

will further investigate the previous research that has been conducted in teacher education in the areas of statistical reasoning, inquiry-based learning, technology use, and issues of equity and assessment, particularly if there is research where any of these regions overlap. Where the research is missing within the realms of teacher education in any of these areas, a broader conceptual research base that was used to inform the dissertation research will be discussed. Chapter 3 will present the theoretical framework driving the study, research design, data collected (both qualitative and quantitative), and methodology used to analyze the data with respect to the research questions. Chapters 4 and 5 will lay out the results of quantitative and qualitative analyses of the data in light of the research questions presented here. Finally, Chapter 6 will discuss these results and make suggestions for their implications for research and practice.

Chapter 2: Review of the Literature

The purpose of this study, as described in Chapter 1, is to examine prospective teachers' use of the concepts of variation and distribution in conducting technology-supported data investigations of equity and fairness in accountability. Examining this particular intersection naturally relies a great deal on previous research on literature in teachers' statistical reasoning and beliefs about equity. But it also requires an understanding of inquiry, assessment, and technology, particularly as they relate to teacher education. This chapter will examine each of these topics, highlighting the issues needed to implement reform and further their research.

2.1 TEACHER EDUCATION

Calls for change in teaching methods have been with us now for decades (see e.g., Dewey, 1938/1997; Lakatos, 1976; Schwab, 1978a; Tyler, 1949/1969), but researchers indicate that little has changed in America's classrooms (Cuban, 1990; Stigler & Hiebert, 1999). And although teachers are the vehicle and target of reform (Wilson & Berne, 1999), professional development of practicing teachers continues to be generally poor despite millions of dollars spent nationally (Cohen & Hill, 2001).

A report by the National Commission on Teaching and America's Future (1996), based on their two-year study of teacher education research, begins with the claim that "the single most important strategy for achieving America's educational goals [is] a blueprint for recruiting, preparing, and supporting excellent teachers in all of America's schools. … A caring, competent, and qualified teacher for every child is the most important ingredient in education reform" (p. 9). This strategy is not disputed, but what is contested is the path to its realization. That is, what sort of preparation and development of teachers is needed to reach this goal? What kinds of knowledge and skills do teachers need to develop in order to be good teachers, particularly at the secondary level? Can anyone with adequate content knowledge teach? What instructional practices must be learned? And what understanding do they need to have of their students, particularly students with backgrounds that differ from their own?

In order to address these questions and examine what is needed to improve teacher education, we need to examine the current state of teaching and identify critical areas in need of change. This section will examine some of the difficulties in the current

state of teaching that has hindered the realization of the vision of reform and some of the levers documented in the literature needed to move the reform forward.

2.1.1 Current State

The National Council of Teachers of Mathematics (NCTM) put out their vision of what mathematics learning would look like if efforts were made to improve content, instruction, and assessment practices (1989; 1991; 1995; 2000). Current instruction in the U.S. is frequently focused on "learning terms and practicing procedures" (p. 41, Stigler & Hiebert, 1999). That is, the majority of time in many American classrooms is spent on teachers demonstrating problems on the board and then having their students practice similar problems. These problems diverge little from the demonstrated examples and remain at a relatively simple level. In addition, students practice many of these rather simple problems instead of working to solve a few, challenging problems. Even problems that are considered applied (i.e., problem solving) frequently follow a restricted and predictable format.

One major problem is that teachers frequently see their role as "simplifying" mathematics by replacing ambiguous contextual mathematical cues in a problem with a "works-every-time" rule. In doing so, they often feel they are making connections easier for their students. Boaler (1997) described very eloquently, however, how this "simplifying" actually makes the mathematics more difficult for the students to learn:

> Teachers gave the students these 'handy hints' or rules to make mathematics questions easier, more straightforward, for students. The teachers understood the mathematics they were talking about and from that base of understanding the rules appeared to be helpful to them. But the students did not understand the rules they were learning or the way that these rules related to the different situations they encountered. They did not locate the rules within a broad mathematical framework and they did not develop a real sense of what they meant. … They view the procedures as abstract rules to be learned and to which they should adhere. Rules may be easy to learn, but difficult to use if they have not been placed within a wider sphere of understanding (p. 26-27).

The NCTM Standards promoted a move away from overly procedural mathematics and towards greater understanding of underlying concepts. The Standards encouraged teachers to focus less on passive transmission of information through lecture and rote practice and more on engaging students through discussion, cooperative learning, the use of technology, and hands-on activities. In addition, they encouraged

teachers to deepen the level of mathematical content and increase the use of authentic problems. In an extensive study of instructional practices in the U.S., Germany, and Japan, Stigler and Hiebert (1999) found that the majority of American teachers felt that the main point of a lesson was to teach students a particular set of procedures and skills. In contrast, German instruction focused on difficult mathematical content and advanced procedures, while Japanese classrooms more commonly were marked by structured problem-solving. Nearly all American teachers were aware of the NCTM Standards and most felt that they were implementing the reforms. However, Stigler and Hiebert's video analysis of American classrooms found that American teachers frequently adopted the trappings of reform (use of cooperative groups, manipulatives, calculators, real-world problems, and writing), but rarely changed their basic approach to teaching mathematics. In fact, Stigler and Hiebert found that "reform teaching, as interpreted by some teachers, might actually be worse than what they were doing previously in their classrooms" (p. 106).

2.1.2 Changes needed to move reform forward

The National Science Education Standards (NSES) assert that "Professional development for teachers of science requires learning essential science content through the perspectives and methods of inquiry" (p. 55, National Research Council, 1996). Although the *Principles and Standards for School Mathematics* (National Council of Teachers of Mathematics, 2000) put out by the National Council of Teachers of Mathematics (NCTM) do not specifically address inquiry, they agree that teachers must learn mathematical content through worthwhile mathematical tasks, engaging in mathematical discourse, and developing an ability to *do* mathematics.

Ball (1996) writes: "If teacher educators aim to prepare teachers who can teach mathematics for understanding, they must create opportunities for teachers to build connections with mathematics, not just as teachers but as learners themselves" (p. 39). In his work with pre-service elementary teachers, Simon (2000) worked to create opportunities for teachers to learn mathematics in an environment characterized by inquiry, with the idea that it "may lead teachers to attempt to create the same experiences for their students" (p. 598).

Several researchers have urged teacher educators to assist teachers in changing practice by immersing them in *doing* content-rich activities and authentic open-ended inquiry (Boaler, 1997; Polman, 2000; Simon, 1995; 2000). Fosnot (1996) concurs that

teachers need experience learning in an environment they themselves are trying to create as teachers, however the need goes beyond learning reform-type pedagogy.

> If understanding the teaching/learning process from a constructivist view is itself constructed, and if teachers tend to teach as they were taught, rather than as they were taught to teach (Jones, 1975), then teacher education needs to begin with these traditional beliefs and subsequently challenge them through activity, reflection, and discourse. … Most importantly, participants need experiences as learners that confront traditional views of teaching and learning in order to enable them to construct a pedagogy that stands in contrast to older, more traditionally held views. (p. 206).

One of the difficulties that NCTM experienced in putting forth its initial vision for reform (National Council of Teachers of Mathematics, 1989) was in the misinterpretation by thousands of teachers of what "activity-based" learning involved (Burrill, 1997b). Science education has experienced similar difficulties. "If, for example, students spend their time making Jell-O molds of dinosaurs and everyone calls the result 'hands-on science', no improvement over the lecture system will have been realized" (p. 24, Powell, 1994). Unfortunately, many teachers focused on the activities and lost sight of the mathematics or science that they were trying to emphasize. If teachers presume that activities, applications, and even symbolic procedures intrinsically carry the intended mathematical connections, then they will feel that there is no real need to make these connections explicit.

Some researchers have documented that lack of content knowledge may be one reason why teachers miss opportunities to help students make these connections. Few would debate that content knowledge is critical for teaching (e.g., Ball, 1996; 2002; Polman, 2000), and there is general agreement that most teachers' understanding of mathematics is not sufficiently deep for them to instill depth of understanding in their students (Ma, 1999). There is much controversy, however, about what kind of mathematical experiences teachers need to have in order to teach, and who should provide it (Fennema & Franke, 1992).

A recent RAND report (Ball, 2002) recommended an agenda for research in mathematics education, and proposed that the development of teacher content knowledge as one of the greatest areas in the field in need of research. One might assume that an advanced understanding of mathematics or science, for example by obtaining a major in their subject, would certainly be sufficient for teachers to have a

good understanding of the content at the secondary level. In Texas, the State Board of Educator Certification (2004) recently enacted a policy that would allow anyone with a bachelor's degree to teach secondary level courses in their major, just by passing a basic test. This poses an important question: Is advanced content knowledge enough? What else is needed to be an effective teacher in mathematics and science?

The RAND study argued that understanding mathematics goes far beyond traditional coursework in mathematics, but also includes development of *mathematical practices* (p. 24, Ball, 2002):

> Noting that expertise in mathematics, as in any field, involves more than knowledge, we propose an explicit focus on mathematical know-how — what mathematicians and mathematics users *do*. We refer to these things they do as *mathematical practices*. Being able to justify claims, using symbolic notation efficiently, and making generalizations are examples of mathematical practices. Such practices are important in both learning and doing mathematics. Their absence can hamper mathematics learning.

Little research, the study reports, has been done on teachers' development of mathematical practices, particularly at the secondary level.

For secondary teachers in mathematics and science, three kinds of content knowledge are critical: (1) an understanding of advanced content (beyond the secondary curriculum), what one would normally gain from a major in the subject; (2) a deep understanding of the secondary-level content (the content teachers teach), particularly since many have not developed a foundational understanding of these topics beyond their own experience as secondary students; and (3) understanding of how to teach the secondary content, or pedagogical content knowledge.

Simon (1995) contends that although the education community generally agrees that moving towards reform-based teaching methods is desirable, researchers have not yet come up with a model of what teaching from a constructivist perspective entails or where teachers should gain the experiences necessary to put reform-based methods into practice. Traditional professional development models are usually provided through workshops, institutes, and coursework. Workshops and institutes, however, are frequently too short to impact teachers' instructional practice positively and have frequently neglected development of teachers' content knowledge (Castro-Filho, 2000; LaChance, 1999). Content area coursework, often conducted in traditional lecture mode, tends to reinforce teachers' reliance on conventional practices. A promising departure

from these forms of professional development are projects focused on reform-based curriculum (implementation, development, adaptation, or through the use of replacement units), immersion experiences and partnerships with scientists and mathematicians, and examining student thinking through scoring assessments (Loucks-Horsley et al, 1998).

Cohen and Hill (2001) reported on a large-scale survey of teachers in California about their professional development and teaching beliefs and practices. Their analysis found very few factors that had a positive influence on teacher practice towards reform-based methods after years of attempts to create change, including expected factors like administrator support for change, and teacher beliefs consistent with the reform. Two elements of professional development appeared to have the greatest influence on teacher practice: teachers who attended extended workshops where teacher learning was focused on either a piece of reform curriculum they were to teach (e.g. replacement units, Marilyn Burns workshops) or where their learning was focused around study of students' work on the (then) state assessment, the California Learning Assessment System (CLAS). In the latter case, the authors theorized that if teachers saw that their students were struggling with problems on the assessment, they would be motivated to teach their students differently. In this case, the assessment (CLAS) was focused on reform-based mathematical content, so this finding is of concern given that across the nation the content tested on current statewide high-stakes tests are contrary to the vision of reform.

One exemplary program of professional development that works to develop teacher learning of practices within their content area is the National Writing Project. The National Writing Project is well known for its focus on developing teachers' *writing practices*. Developed in 1974, the National Writing Project (Gray, 2000; National Writing Project, 2002a) works with teachers to develop their own writing ability. The premise of the project is that if teachers better understand the process of writing through learning to be better writers, this in turn will improve the way that they teach writing. The focus on teachers' *doing* writing, rather than learning how to *teach* writing is unique and highly effective (Lieberman & Wood, 2003; National Writing Project, 2002b, 2002c).

What is promising about the National Writing Project is its simultaneous focus on the development of teacher's own content knowledge and opportunity to be learners

in an environment that models one they are being encouraged to use in their own classrooms. The dissertation aims to study preservice teachers' understanding of a particular domain of content, statistics, in the process of conducting inquiry. These two areas of research are the focus of the next two sections.

2.2 STATISTICAL REASONING

A basic assumption in the accountability system is that teachers will use the results of their students' test scores to guide and improve their instruction. A poor understanding of data and statistics, however, can lead to stereotyping if teachers are not aware of natural variation or do not attend to distributions of scores (Confrey & Makar, in press; Wiliam, 2003). But statistics is a content area in which even mathematics and science teachers often have little background. One of the main goals of the course was to improve the preservice teachers' content knowledge in statistics so to give them the skills and concepts needed to interpret data from student assessments, as well as to improve their content knowledge of statistics for teaching. This section will examine the literature used to inform both the study and the course in the area of teachers' statistical reasoning. Central to the study is preservice teachers' understanding of the particular concepts of variation and distribution. These two particular areas of research—teachers' statistical reasoning, and reasoning about variation and distribution—are only now beginning to develop. Most of the literature on statistical reasoning has focused on children; the literature on children's statistical reasoning is useful here, as some studies have documented (e.g., Confrey & Makar, 2002; Rubin, 2003) that in areas where adults are novices, they tend to go through similar process of development. This study aims to add to the research of these topics, particularly how prospective teachers conceptualize and apply their understanding of variation and distribution.

Statistics is a fairly new topic of study in schools. Initially, statistics courses were taken primarily at the university level for research methods (Utts, 2002), but the prominence of data in the information age has prompted a call for a statistically literate citizenry (National Council of Teachers of Mathematics, 2000; National Research Council, 2000) and pushed probability and statistics into the curriculum as early as kindergarten and first grade (National Council of Teachers of Mathematics, 1989; 2000; TERC, 1998; Texas Education Agency, 1997). This has had an effect of shifting the research in statistics from a focus on probabilistic reasoning about of randomness and

heuristics in the mid to late 1900s, to recent research focused on statistical reasoning about concepts in the school curriculum.

Some research has shown that children's thinking about randomness is initially deterministic in nature (Lehrer & Schauble, 2000a; Marshall, Makar, & Kazak, 2002) but that ideas of randomness can develop through experience and instruction (Batanero & Serrano, 1999). Fischbein and Schnarch (1997) found that misconceptions decreased with age in problems that were clearly probabilistic in nature (e.g. outcome of tossing a die). Pfannkuch and Brown (1996), however, reported that even with statistical training university students in their study had difficulty letting go of strong intuitive beliefs and frequently dealt with the conflict by holding dual beliefs. Several researchers have formulated frameworks within concrete-symbolic reasoning to describe children's probabilistic thinking and reasoning about data (Jones et al., 1999; Biggs & Collis, 1982; Friel, Curcio, & Bright, 2001; Langrall & Mooney, 2002; Shaughnessy, 1992).

Until recently, research on children's thinking reported in the misconceptions literature could be perceived as demonstrating deficit thinking about children rather than considering that "one must seek to model their problematic and not presume it is identical to one's own" (p.137, Confrey, 1991). Other researchers have reported that the difficulty students experience in understanding randomness might originate not in shortcomings of student thinking, but could be a result of a negative effect of the domination of rule-based instruction in school (Duckworth, 1996; Kamii, 1994; Shaughnessy, 1992) where the approach is based more on *mathematical reasoning* than *statistical reasoning* (delMas, in press; Shaughnessy & Bergman, 1993; Stuart, 1995). Although schools often feel a push to move students to use conventional bar graphs as early as first grade (National Council of Teachers of Mathematics, 2000; Texas Education Agency, 1997), doing so may lead to a recipe approach to reasoning with data with the treatment of data as just numbers, lacking context and practical importance (Konold & Higgins, 2002). Confrey & Smith (1995) argue if children's classroom experiences stress addition and counting at the expense of other mathematics, it may persuade students to believe that these are the only choices for handling numerical information. This result extends to older students as well, where researchers studying university students' knowledge of significance testing (Gardner & Hudson, 1999) found that students often applied tests without an understanding of their purpose. Other researchers have warned that the premature use of hypothesis tests at the high

school and undergraduate level can aggravate the black and white misuse of the accept-reject dichotomy of statistical tests (Abelson, 1995; Reichardt & Gollob, 1997).

2.2.1 Moving the field forward to improve instruction—from documenting misconceptions towards developing conceptual reasoning

Shaughnessy (1992) in his landmark synthesis and state of the literature in probability and statistics reported that there appeared to be three critical barriers to improvement of statistical teaching and learning: the lack of statistics in the curriculum, poor teacher background in statistical conceptions and content, and poor intuition about stochastics among teachers and students in general. Shaughnessy generated the following list of areas of research specific to stochastics teaching and learning needed to move the field forward (p. 489-490): (1) development of standard assessment instruments on student conceptions of probability and statistics; (2) research on secondary students' conceptions; (3) cross-cultural studies; (4) research on teachers' conceptions; (5) teaching experiments documenting the effects of instruction on stochastics learning; (6) research on the effects of computer software on learning; and (7) the role of metacognition on decisions under uncertainty.

Since Shaughnessy's (1992) report, statistics education has taken a tremendous shift in its research emphasis from studies of probability conceptions (although still present) towards school-based curriculum issues: children's concepts of average (Mokros & Russell, 1995; Watson & Moritz, 2000), graphical representation (Ben-Zvi & Arcavi, 2001; Friel et al., 2001; Lehrer & Schauble, 2000a, 2000b; Moritz, in press; Roth & McGinn, 1997), technological tools (Bakker, 2002; Ben-Zvi, 2000; Biehler, 1997; Burrill, 1997a; Garfield & Burrill, 1997; Konold, 2002a), and assessment issues (delMas, Garfield, & Chance, 2001; Gal & Garfield, 1997; Garfield & Chance, 2000; Konold & Khalil, 2003). Other recent research focuses on children's conceptions of sampling (Shaughnessy, Watson, Moritz, & Reading, 1999; Watson, 2002) and inferential reasoning (Watson & Moritz, 1999).

Besides a transformation in content, another shift in the literature in statistical reasoning has been from a focus on individual thinking in clinical interviews towards an emphasis on developing and studying statistical reasoning within a classroom community. These studies, with a more sociocultural perspective, have stressed the need to go beyond research on statistical concepts and procedures towards children's development of statistical inquiry and data modeling in a collaborative setting (Cobb,

1999; Hancock, Kaput, & Goldsmith, 1992; Konold & Higgins, 2002; Lehrer & Schauble, 2000a, 2000b).

2.2.2 From mastering content towards developing a mindset of inquiry

Statisticians have often noted that even students who can perform statistical procedures correctly frequently lack the ability to think statistically. Pfannkuch and Wild (2001) claim that part of the problem is that 'statistical thinking' has remained undefined. Snee (1990) defined statistical thinking from the perspective of the quality control industry:

> I define statistical thinking as *thought processes*, which recognize that variation is all around us and present in everything we do, all work is a series of interconnected processes, and identifying, characterizing, quantifying, controlling, and reducing variation provide opportunities for improvement. … The importance of statistical thinking derives from the fundamental principle of quality put forth by W. Edwards Deming: Reduce variation and you improve quality (p. 118).

According to Garfield and Gal (1999), statistical reasoning includes reasoning about data, representations of data, statistical measures, uncertainty, sampling, and association. Wild and Pfannkuch (1998; 1999) put forth that the development of a mindset of probing, evaluating, and describing, an awareness of contextual constraints involved, and a balance of curiosity and skepticism are critical to applying statistics and these authors have developed an initial framework to describe statistical thinking.

Lehrer and Schauble (2000a) argue the importance of meaningful instruction in developing this mindset, that statistical "reasoning seems to be mastered only over an extended period and depends on thoughtful instructional support and repeated opportunities for practice and use" (p.114). Data-based inquiry has received a lot of attention at the elementary school level where researchers note the difficulty students have in creating measurable conjectures and determining appropriate data (e.g., Hancock et al., 1992) as well as relating their findings back to their original research questions (Konold & Higgins, 2002; Marshall et al., 2002). Shaughnessy (1992) asserts the premise that "data-modeling activities, such as constructing, manipulating, and interpreting data, consist of skills and concepts that come before the process of actually doing statistics" (p. 484). It is important not to lose sight of the *interaction* of content knowledge with context. As a study of experts' approach to ill-structured problems

indicated, inquiry "will lead to inadequate solutions unless the individual has and employs substantial knowledge of the domain" (p. 283, Voss & Post, 1988).

The literature on content knowledge of teachers has strongly indicated that there are deficits in teacher knowledge of the subject matter that they teach (Fennema & Franke, 1992; Ma, 1999). Few of these studies, however, document and value teachers' less formal conceptual understanding, particularly in the domain of stochastics. Researchers have stressed that intuition and mindset about data and uncertainty are critical elements of statistical thinking that are systematically ignored in education (Moore, 1997; Wild & Pfannkuch, 1999; Snee, 1990). This neglect is made more serious by the overwhelming prevalence of statistical concepts being taught in schools as part of the mathematics curriculum, where the dominating emphasis has been on articulating and examining concepts in which uncertainty plays no role and requiring a very different kind of thinking than what is needed for statistical thinking (delMas, in press).

> The incompatibility between stochastics and traditional school mathematics results basically from the fact that these two fields of knowledge result in *opposed types* of knowledge and development of knowledge in the classroom. This difference can be characterized as follows; while teachers in traditional school mathematics have an understanding of the subject which is based on a hierarchically ordered and cumulatively constructed stock of knowledge, which is taught in a linear sequence and learned step by step, in the case of statistics and probability, they are always confronted from the outset with a complex multitude of mathematical, situation-specific and interpretative aspects which indicate that stochastical knowledge has a complex systemic structure and cannot be learned in a linear sequence, but only in a holistic mode (p. 7, Steinbring, 1990).

Although the traditional practice of teaching mathematics as a subject of deterministic and hierarchically-structured knowledge is contrary to the vision of instruction put forth by the National (U.S.) Council of Teachers of Mathematics *Principles and Standards for School Mathematics* (NCTM, 2000), the process of change has been slow and the instructional core has remained relatively unchanged despite attempts to reform it (Cuban, 1990; Cohen and Hill, 2001). The difficulty in changing mathematics instructional practice in schools provides some hint to the kind of struggles the statistics education community expects in reforming and deepening statistics instruction in schools. This highlights the importance of changing the mindset of preservice teachers about mathematics and statistics before they begin their careers.

Certainly if we are to move students' away from isolated procedural knowledge in statistics towards an ability to manage a complex and uncertain world, teachers must develop experience in the use of data and statistics towards this perspective.

2.2.3 Teachers' experience with statistical inquiry—the pilot study

Critical for this dissertation study is an understanding of how teachers' reasoning of statistical concepts plays out in the context of authentic statistical inquiry. Although little research on teachers in this area has yet been conducted, much can be gained by examining the research on children's learning of statistical inquiry. At the school level, several researchers have noted that students often focus on personal factors and individual points in a data set (Bakker, 2001; Hancock et al., 1992; Konold & Higgins, 2002; Marshall et al., 2002). Although it may seem that student understanding must be very different than teachers' understanding of data, Confrey and Makar (2002) found, in a pilot study for this dissertation, that teachers may also initially perceive data as individual points rather than as a distribution. This indicates that teachers without statistical training may think more similarly to their students than previously assumed.

The pilot study involved a six-month partnership with an urban middle school to conduct a professional development series designed to improve teachers' understanding of their students' TAAS results (Confrey & Makar, 2002; Confrey, Makar et al., 2001; Makar & Confrey, 2002; in press). Our research group found that through experience in data investigations and conducting their own inquiries into student test data, the middle school teachers in the project not only gained statistical understanding, as measured on the pre- and posttest, but were better able to develop more statistical habits of mind that parallel those found in research on student data inquiries (Bakker, 2004; Hancock et al., 1992; Konold & Higgins, 2002; Konold, Higgins, Russell, & Khalil, 2003; Marshall et al., 2002). For example, we saw teachers move from a focus on individual data points and linking these to their knowledge of specific students, to an attention to global trend, variation, and distribution (Confrey & Makar, 2002). Confrey and Makar cautioned, however, that when more complex concepts, like sampling distributions, were not adequately constructed *by the teachers*, they can use statistical tools mechanically, without carefully examining their relationship to the data (Makar & Confrey, in press). This corroborated the findings of other researchers (Chance, Garfield, & delMas, 2001; Saldanha & Thompson, 2001) that sampling distributions are very difficult conceptually. Biehler (1997) reports similar difficulties in preservice teachers'

distinguishing between data at the group and individual level when comparing distributions. He notes that preservice teachers deeply struggle with understanding concepts of distribution more generally and that technological tools can both promote their understanding and become an obstacle. Some of their difficulty in understanding distributions he attributes to the lack of language we have in describing distributions beyond the level of statistical summaries. "This difficulty may not be surprising because data distributions are usually not characterized as concepts in courses of elementary data analysis. Distributions are emphasized in probability theory but in an entirely different context that students find difficult to apply in data analysis" (p. 180).

The pilot study also found that setting data within an authentic context relevant to the teachers motivated them to articulate richer descriptions, including attention to variation, when comparing groups than when given the same data without a context. Finally, conjectures that the teachers created at the beginning of the study focused on simple well-defined questions that would produce yes or no answers, or ones that were too complex to measure with available data. By the end of the pilot study, the teachers' inquiry projects displayed a much deeper level of thinking about what is possible to uncover in data. Examples of teacher inquiries included: (1) the contrasting trajectory of remediation in large urban districts to other districts in the state; (2) the effectiveness of practice tests in predicting performance and assisting students most at risk of failing the state exam; and (3) by comparing the performance of "average" (middle 50%) students on the 7^{th} grade TAAS and these same students on the Iowa Test of Basic Skills, one preservice teacher investigated whether the TAAS was testing students at their respective grade level. This was a long way from initial questions they proposed to investigate at the beginning of the study, such as: On which test objective do students perform the worst?

What was learned in the pilot study was that teachers need extended opportunities to develop their own inquiries and that given these opportunities, their understanding of the complexity of the context under study is deepened; in our case, the teachers developed rich interpretations of their students' results and designed sophisticated inquiries to investigate complex questions of interest to them. Furthermore, the teachers improved their content knowledge in statistics and became comfortable using dynamic learning-oriented software. In addition, by deepening their understanding of the context through their data-based inquiries, there was some

evidence that the teachers developed a greater sense of understanding about the need to nurture all of their students.

2.2.4 Moving statistics towards a conceptual focus on variation and distribution

It is generally agreed in the statistics education community that there is a need to develop instructional methods to assist students to develop a more holistic view of distributions, and recent research has turned in this direction (e.g., Bakker, 2001; Ben-Zvi & Arcavi, 2001; Konold & Pollatsek, 2002; Lehrer & Schauble, 2000a, 2000b, 2002; McClain & Cobb, 2001). In addition, Shaughnessy, Watson, Moritz, and Reading (1999) have noted the over-emphasis in the school curriculum and standardized tests on measures of center and have pushed for more focus on variation and sampling. Meletiou (2003) urged researchers to focus more on variation by compiling and publishing a list of previous research on various aspects of variation. The third and most recent International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-3) was completely dedicated to research on reasoning about variation (Gal, in press; Garfield, Ben-Zvi, & Mickelson, 2002; Lee, 2003), and the fourth SRTL in 2005 will focus on reasoning about distributions (Pfannkuch, 2004).

With the development of technological tools, recent work with elementary and middle school students has shown promising results in reversing the momentum of statistics as formula-based towards creating an environment that encourages students to develop a distribution-perspective of data. For example, Cobb (1999) has used his *minitools* to ask students to compare the performance of two brands of batteries using non-standard representations. Through this work, we see how the tools and the task work together to encourage students to find a need for variation to describe the relative reliability of the batteries—an important consideration in performance.

The recent focus on children's conceptions of variation and distribution has also brought about a revolution towards respecting children's non-standard language and development of informal, intuitive conceptions about distributions. Bakker (2001, 2004, in press) found that with Cobb's minitools and an innovative learning trajectory, he was able to urge his middle school students to think about variation and distribution in arguing their case with the "bump" of a mound-shaped distribution. Konold and his colleagues (2002) termed this 'bump' a *modal clump* in his work with middle school students, emphasizing that in problem solving with data, the middle hump of a distribution was a frequently identified portion of a mound-shaped distribution. Konold

(2002b) and his colleagues also noted that students tend to divide distributions into three categories: low, middle, and high. As a result, he developed a 'hat plot' in the software *Tinkerplots* which allows students to isolate these three groups based on visual or quantitative methods. *Tinkerplots* (Konold & Miller, 2002), a statistical software package under development in the same spirit as the minitools and other innovative technologies, helps students explore data (represented individually as large dots, bars, pie-slices, or icons) with primitive actions of sorting, sequencing, and stacking. This enables students to create their own data representations and construct rather than accept standard graphical forms such as scatterplots (Konold, 2002a), boxplots, and histograms, or, create their own graphical representation that is meaningful to them. We do not mean to infer that the use of technology will lead students to construct a deeper understanding of variation and distribution. From the 1996 Roundtable conference on the use of technology in learning and teaching statistics, Burrill (p. 71, 1996) summarizes one group's discussion:

> A good teacher knows how to not just "use" technology but to make it an effective part of the teaching and learning process. The group was reminded that probability and statistics requires a different sort of thinking and that many teachers do not think in probabilistic and statistical terms. … Both the need for the professional development of teachers in order for them to effectively teach statistics and use technology to teach statistics, as well as the need to have a clearly defined statistics curriculum were underlying themes of the rest of the discussion.

With or without the influence of technology, less is known about how adults construct basic concepts of variation and distribution. The notion of teachers' mindset and basic conceptions about data has been of great concern in the research community for over a decade (Hawkins, 1990; Shaughnessy, 1992), but despite this, Shaughnessy (2001) reports that he is "not aware of any research studies that have dealt specifically with teachers' conceptions of variability, although in our work teaching statistics courses for middle school and secondary school mathematics teachers we have evidence that many teachers have a knowledge gap about the role and importance of variability." Noss, Pozzi, and Hoyles (1999) examined nurses' interpretation of "average" blood pressure in time-series data and found that their perception of average in the context of their professional work was highly dependent on their interpretation of the variation in the data.

2.2.5 Agendas and future trends in research

At the secondary and university levels, researchers have begun to set an agenda for critical learning and challenges in teaching and learning statistics beyond middle school (Scheaffer, Watkins, & Landwehr, 1998; Utts, 2002), including an understanding of correlation verses causation, importance of sample size in inference, sources of bias, understanding distribution and variability, experiments verses observational studies, formulating hypotheses and conjectures, and inferential reasoning. Little research on these elements has been conducted for this population (Meletiou, 2000).

Batanero, Garfield, Ottaviani, and Truran (2000) set out research goals for the statistics education. They proposed research in models to understand the evolution of statistical reasoning, the development or extension of current learning theories to support the teaching-learning of statistics, documentation of activities and learning environments that help students construct deep understanding of statistical concepts, and studies which document the effect of technological tools on student learning. In addition, they noted that despite Shaughnessy's 1992 paper, research on teachers' conceptions of probability and statistics had still not been sufficiently researched, and they urged the research community to examine effective ways to train current and future teachers of statistics. Recent research on teachers' understanding of statistics has included case studies of elementary teachers (Heaton & Mickelson, 2002; Mickelson & Heaton, in press), studies of preservice teachers (Begg & Edwards, 1999; Canada, 2004; Edwards, 1996; Makar & Confrey, 2003, 2004, under review), and work with practicing teachers (Confrey & Makar, 2002; Makar & Confrey, in press; Rubin, 2002).

The literature in statistical reasoning indicates that there is growing interest in how students and teachers come to develop and use statistical concepts in the context of conducting inquiries. I now turn to developments in research in this area. Much of the literature on inquiry-based reasoning lies outside of statistics in science education, philosophy of science, and broader domains.

2.3 INQUIRY

At the heart of statistical reasoning is the ability to investigate: to formulate and test hunches, justify conjectures with evidence, and communicate findings, if any, with a convincing argument. Fundamental beliefs about statistical reasoning are echoed beyond the domain of statistics to calls for inquiry-based teaching and learning in math

and science. Inquiry is arguably central to scientific study. The National Science Education Standards calls for teachers to provide inquiry-based learning opportunities for their students to help them to learn how to *do* science, and recommends that teachers assess students' ability to perform an inquiry (National Research Council, 2000). This call uses the term 'inquiry' in two ways. It refers to the teaching and learning strategies that employ investigations to enable mastery of scientific concepts; additionally, it refers to the skills, conceptions, and mindset that students should develop to be able to design, conduct, and understand the nature of scientific investigations (p. xv).

Although the National Council of Teachers of Mathematics (1989; 2000) is less explicit about urging teachers to have students conduct mathematical inquiries, the principles they call for teachers to use when teaching mathematics are consistent with inquiry-based learning: formulating questions, making and investigating mathematical conjectures, developing mathematical arguments, evaluating predictions based on data, building mathematical knowledge through problem-solving, and communicating mathematical thinking clearly to others. Inquiry is not unique to science and mathematics, but is part of a larger goal of curricular reform in education towards 'hard content', active learning, and more authentic achievement, with an emphasis "on students learning to produce knowledge, rather than simply reproduce knowledge. Authentic achievement, then, requires disciplined inquiry: the use of prior knowledge, in-depth understanding, and the integration of ideas and information" (p. 11, Porter, Archbald, & Tyree, 1990).

Mathematical thinking encompasses more than what is often presented in school mathematics, i.e., a perspective of mathematics as a set of fixed rules and relationships to be learned and applied. This view of mathematics, Thompson (1992) reports, is often held by teachers and is accompanied by the belief that mathematics is *a priori* true. An inquiry epistemology challenges this commonly held belief and instead provides alternatives: mathematical knowledge is fallible; mathematical knowledge is created through a non-linear process which often includes the generation of multiple hypotheses; the production of mathematical knowledge is a social process; negotiation is a critical determining factor of the truth value of mathematical knowledge (p. 205, Siegel & Borasi, 1994). An inquiry-based view includes many other aspects of the essence of the discipline including: how ideas are communicated, what counts as evidence, methods and processes used, and "*lenses* through which we look at the world

and interpret it" (p. 18, Mansilla, Miller, & Gardner, 2000). This perspective is more philosophical, one that broadens the discipline of mathematics from a network of knowledge to a system of beliefs and perspectives.

Generating inquiry is a skill that requires breaking through a mindset of having to find an answer to every question (Wild & Pfannkuch, 1998). Schwab (1978a) argues that critical to inquiry in education is a development of a 'polyfocal conspectus'—a pluralistic view of theory, or a belief that multiple theories can provide multiple insights into practical problems. Through an understanding of complexity and a polyfocal conspectus,

> students would at least be saved from the expectation, forced on them by earlier doctrinaire education, of a unique solution to every problem. … Students might begin to discern the fact that the members of a plurality of theories are not so much *equally* right and *equally* deserving of respect, as right in different ways about different kinds of answers to different questions about the subject and as deserving different respects for different insights they are able to afford us (p. 338-339).

Complexity and uncertainty are key elements to developing inquiry. The complexity of a problem provides maximum freedom in an inquiry and provides a prime material for demonstrating multiple perspectives, according to Schwab (1978a), although uncertainty provides a critical motivator. Peirce (1923/1998) recognized that inquiry is embedded in a need to resolve doubt. It is when students internalize a problem so that the problem becomes their own that they feel a need to seek resolution. Confrey (1991) calls this internalization of a problem the *problematic*, when students make a problem into their own, seeing it from their own perspective. Dewey also stated the importance of creating a problematic in his statements about inquiry; that in investigating, the inquirer needs to feel disequilibrium in order to push ahead to reach a resolution and that an internal cognitive motivator kicks in (Hickman, 1990). Inquiry-based learning, however, is not akin to *discovery learning*—where students are left to 'discover' mathematics on their own—but rather requires a great deal of assistance and support from teachers. Confrey (1991) emphasizes that during inquiry students are not aiming to simply uncover what the teacher wants them to find, or to recreate a particular process, but to reconceptualize an idea based on their own "problematic".

Math and science in an inquiry-based classroom is complex, collaborative, generative, student-centered, and no longer dominated by a transmission mode of

teaching (National Research Council, 2000; Siegel & Borasi, 1994). Creating an inquiry-based classroom environment requires that teachers help students develop a community of practice (Cobb, 1999). Koedinger (1998), working in the domain of geometry, argues that the ability to investigate, generate and test 'interesting' conjectures, and produce a convincing argument, can in fact be learned by anyone, but "requires self-reflection and extra cognitive resources that come only from practice" (p. 333). Schwab (1978a) contends that inquiry can be facilitated by (1) providing a basis of knowledge and terminology; (2) demonstrating how complex problems lead us to multiple questions; and (3) showing how each of these questions can lead to a separate inquiry. By creating initial structure for students to conduct inquiry, the teacher is able to provide students a first glimpse of a polyfocal conspectus. Schwab goes on to describe how cycles of inquiry are developed in a classroom, with each cycle beginning with a basic grounding in background knowledge and relevant theory, which are then applied and challenged through examination of initially structured case studies and increasingly autonomous inquiries.

Conducting an inquiry as a learner is an area in which teachers have little experience; for them, the process of change is difficult (Cuban, 1990). Helping teachers to reconstruct their view of classroom instruction is one of the greatest challenges for those who teach prospective teachers in math and science (National Council of Teachers of Mathematics, 2000; National Research Council, 2000). Technology, which plays an important role in the course used in this study, is often considered to be a potential and powerful impetus for changing instruction in schools. The research basis of the potentials and obstacles of using technology to forward reform is the topic of the next section.

2.4 TECHNOLOGY

Although not the central focus of this dissertation, technology plays a key role in the study in two ways. For one, the particular statistical software, Fathom (Finzer, 2001), was chosen because of its potential as a critical tool underlying the prospective teachers' inquiries by enabling them to visualize relationships in the data, and conduct the analyses underlying their investigations. Secondly, the software chosen for the study was specifically developed for *learners* of statistics; by becoming facile with the software, the idea is that the prospective teachers will begin to use it and others like it as part of instruction with their own students. The study and treatment were designed to

include these two aspects of technology in order to both support the way that technology can enable the vision of educational reform, and work to counteract the roadblocks that have been documented. How is the use of learning technologies critical to reform, and what barriers to this vision exist? This section will examine the literature for insight into this question by first articulating the vision of technology towards educational reform, then examining the forces that enable and restrict teachers in their use of technologies that support learning. The elements in the literature discussed here, to support the vision and work to counter the barriers, were used in the design of the study.

2.4.1 The vision

Technology has the possibility of influencing changes in curriculum, instruction, and assessment (National Research Council, 2001b). Although most states are now requiring that new teachers be able to use 'appropriate' technology in order to be certified (e.g., California Commission on Teacher Credentialing, 2000), the *kind* of technology that teachers employ in their classrooms can be crucial to the nature of learning they engender in students (National Council of Teachers of Mathematics, 2000). One potential power of new technologies is the way that it re-energizes the education community to rethink old questions about educational goals, pedagogical practices, the nature of the subject matter, the nature of learning and learners, epistemological relationships between knowledge and the knower, the social structure and controlling powers in the classroom, and the organization of schools (Kaput, 1992).

The opportunities inspired by new technologies have spurred some organizations to develop standards and visions for ways in which technology can work to improve teaching and learning. For example, the Texas State Board of Education (2000) developed their Vision of Technology in Education 2010 that envisioned the possibilities that technology could bring to the education. To meet the vision, they developed a Long-Term Plan for Technology, 1996-2010 that included two-year and six-year goals for students and teachers as well as goals that included parents and community in the final years and responsibilities from teacher preparation programs, professional development, administration, and infrastructure (which would require funding assistance from the State Legislature). Hawkins (1997) added to the vision research on technology use and equitable access to technology. The International Society for Technology in Education (2000) developed its *Technology Standards for*

Students that assist teachers in developing grade-level appropriate technology skills, experiences, and behaviors for their students.

2.4.2 The obstacles

Unfortunately, technology is often considered to be a panacea of a solution to the woes of education and many organizations have jumped on the bandwagon to create visions and standards to improve the use of technology in education even without funding the support and resources that are critical to their implementation. In order to realize the visions such as those above, three challenges need to be addressed: sufficient financial commitment at the local, state, and national levels; support and time for teachers to learn and implement technologies, including training for preservice teachers; and cooperation from software developers to create better tools for student learning, particularly content software for secondary students (Glennan & Melmed, 2000). Hawkins (1997) also includes the challenges of changing the structure of the learning environment and the disconnect between research and policy, including the nature of policy changes. These challenges require commitments that frequently are not aligned with natural political and market forces.

Computers have increased their presence far more quickly than imagined. For example, in 1981 only 18% of schools had computers with an average of 125 students per computer, but by 1993 that number dropped to 14 students per computer with 99% of schools reporting that they had computers (Tyack & Cuban, 1995). As might be expected, however, the presence of machines doesn't indicate how or if they are used in instruction. Although 98% of schools in 1993 reported that they used computers for instructional purposes, there is strong evidence that few teachers actually do and the majority of those who do use them for glorified worksheets, diagnostic tools, or practice of basic skills. In other words, teachers adapt their instructional use to include computers not to change what or how their students learn, but to do the same things they did before (Papert, 1990). One might compare early uses of the computer in the 1980s to the days of the Model T, where "one had to attend at least as much to the vehicle and its operation as one attended to where one was traveling" (p. 516, Kaput, 1992). Since then, much of the interface has become standardized across manufacturers (for both cars and computers) and much of the intelligence needed to operate them has shifted from user to machine.

It is not just the type of technology teachers use that limit its use as a learning tool. While still in its infancy in classrooms, Kaput (1992) warned that limitations in use are less likely to be a factor of the technology "than a result of limited human imagination and the constraints of old habits and social structures" (p. 515). Many teachers who are uncomfortable with new technologies use them to "simply transfer the traditional curriculum from print to computer screen" (p. 516) with computer use that resembles traditional worksheets and structured learning environments and assist teachers in doing their regular work more efficiently, rather than working to transform his or her practice (Tyack & Cuban, 1995).

Learning new technology was an impetus for many teachers to join CoVis (now LeTUS), a research project at Northwestern University, focused on improving science educators' teaching practice through employing technologies. Polman (2000), a researcher at CoVis, admitted that he had difficulty getting his research subject, master teacher Rory Wagner, to use technology with his students except for accessing news and for communication. One of the original tenets of the CoVis project was that technology could be used as a tool to help teachers move towards change in pedagogy; they found, however, that in reality teacher change through using new technology was much more complex. Many teachers use technology as a reward for students rather than as a part of instruction. Polman argues for the continued attempts to use technology to get at teacher change: "In contrast, computers and networking, the latest technologies to be heralded as revolutionizing instruction, have consistently been linked to reforms toward child-centered instruction" (p. 33).

Becker (2000), in his study of teachers' use of the internet, found that the three major predictors for internet use were: (1) teachers' level of classroom connectivity—ranging from no classroom connection to direct connection of four or more computers; (2) teachers' non-internet computer expertise, which was the strongest predictor in teachers' statement of valuing technology as essential for good teaching; and (3) teachers' pedagogical beliefs and practices—that is, whether their instructional practices followed a constructivist-compatible or more traditional, skills-based teaching practice. This study points to fact that although the presence of computers and connectivity are important, teachers' use of technology is strongly influenced by their facility with technology and beliefs about teaching and learning. Therefore, in designing professional

development for teachers, one must attend to both teachers' expertise in using technology and their philosophy of learning.

2.4.3 Equity and technology

The review of the literature that documents the vision and challenges of technology for enabling educational reform would remiss without mention of the digital divide. Technology access and use were reported above, but what was missing was a discussion of *who* had access to technology and differences in *how* it was being used between Whites and Blacks, and between poor and wealthy communities. For example, according to Tapscott (2000), Blacks were two-thirds less likely to have a PC in the home than Whites. This gap is growing wider over time. Similarly, 24% of those without a high-school diploma are interested in using the computer to obtain product information versus 64% of college graduates. In 1997, hardly any (less than 10%) children in lower-income households had a computer at home compared to 70% of higher income children. This gap extends to schools as well. The Clinton administration pledged $500 million so that every school would have internet access. However, with 84,000 public schools in the United States, this meant that the government was willing to fund only 12% of the $50,000 cost of wiring each school. This has the effect of widening the gap even further by making it much less likely for schools in low-income areas to take advantage of this government assistance than those in high-income areas. The gap extends to gender as well (Kirkpatrick & Cuban, 2000). The gap in achievement and attitude towards computers between genders is small at the elementary level, but grows dramatically as children progress through school. By the graduate level, men were six times more likely than women in 1991 to earn a doctorate in computer science. Achievement and attitudes between genders were found to be similar, however, for males and females who had the same amounts and types of experiences on computers.

Not only access, but evidence of *how* computers are being used also differs greatly for different student subgroups. For example, the Office of Technology Assessment (U.S. Congress Office of Technology Assessment, 1988) reported that low-achieving students are less likely than other students to use the computers for high-level reasoning tasks and problem solving, and more likely to use them for drill and practice. Technology access and type of use are critical for improving equity if you consider, that

"who will prosper and who will not will be largely a matter of who is able to enter the computer future of learning" (p. 2, Papert, 1990).

2.4.4 Integrating powerful learning technologies

Bakker (2002) defines two types of software programs that are generally found in mathematics classrooms: *route-type* and *landscape-type*. *Route-type* software facilitates or replaces teacher-directed lessons by guiding students through a specific lesson, or perhaps are more open-ended, but satisfy a single, structured purpose. Their output and the lessons learned are fairly predictable and this provides the teacher with confidence that they will know the outcome, in advance, of student work. Not all route-type software has this diagnostic or worksheet-like purpose, but as discussed above, this has been found to be the overwhelming majority of use in the classroom. Bakker denotes more powerful and open-ended learning software types as *landscape-type*. Landscape-type learning software—for example, Geometer's Sketchpad (Jackiw, 2001), Tabletop (Hancock, 1995), and Function Probe (Confrey, 2000)—give students power to take multiple routes to investigate shapes, data, and functions. These packages are well utilized by project-based or inquiry-based classrooms for their open-ended purpose and their power for exploring "what if" questions with multiple perspectives.

Tinker (1996) put forth four levels of technology integration that indicate the impact technology use is having in the curriculum:

- *Level 1: Substitution*. Technology used to accomplish current curricular goals, but perhaps to assist in obtaining higher levels of comprehension. For example, the use of computers with probes to conduct science labs already in the curriculum.
- *Level 2: Addition*. Technology used to achieve new curriculum goals, perhaps by adding new material to an existing course. For example, international collaboration linked through telecommunications with the TERC Global Lab project.
- *Level 3: Disciplinary Restructuring*. Technology used to redesign course within the discipline. For example, the use of graphing much earlier in the mathematics sequence.
- *Level 4: Interdisciplinary Restructuring*. Technology supports the redesign of courses across the disciplines. For example, the capacity developed by learning systems modeling in a ninth grade mathematics course and built upon in subsequent science courses could allow a broader range of science material at a deeper level.

These levels are very realistic and at the same time challenging. For example, the last level presumes the current structure of education, but is very difficult to implement without large-scale consensus. Tinker adds that many of the curriculum standards for technology use developed for mathematics and science education (e.g. NSES, NCTM, AAAS) are at Level 2, using technology to make increased uses of inquiry in the curriculum, with the mathematics standards beginning to venture into Level 3 by advocating for earlier introduction of topics such as graphing, data analysis, and material that makes use of the computational capacity of technology. Because these standards are discipline-based, it is unlikely they can be relied on to move to Level 4, which would require collaboration between the disciplines.

2.4.5 Research on technology in teacher education

Loucks-Horsley, Hewson, Love, and Stiles (1998) contend that technology can provide teachers with diverse learning experiences, individual growth, and support in developing teacher communities, all critical to professional development. The most common form of technology to assist with these goals has been through the use of email and the Internet. Technology-rich learning experiences for teachers must be developed beyond email and the web, however. NCTM (2000) urges teachers to use appropriate software to support student learning and problem solving through the use of graphing utilities, spreadsheets, dynamic geometry software, and microworlds. Ball (2002), in recommendations for research in mathematics education, adds the use of modeling software, graphing calculators, and computer algebra systems to this list and advocate its thoughtful use to support learning in algebra, particularly algebraic representations. Little is known, however, about how algebra teachers use technology and other materials in instruction—an area of research argued as critical to large-scale change. Several researchers have also pointed to a need to better understand the way that technology changes learners' conceptions of content and strategies for problem solving (e.g., Garfield & Burrill, 1997; Shaughnessy, 1992). In statistics, the content area that is a focal point of this study, researchers have just begun to examine technology use, and note a greater attention given to decisions based on visualizing data representations as a result of the use of software (e.g., Biehler, 1997; 2001; Rubin, 2002).

Further attention to the impact of accountability and high-stakes testing on teachers' use of technology also is in need of further research. For, "the current assessment system, if it relies heavily on standardized achievement tests, can also be a

barrier to experimentation with new technologies because teachers are not sure whether the results they are seeking will be reflected in improved student test scores" (p. 18, U.S. Congress Office of Technology Assessment, 1995). Not just in technology, but more generally, the current accountability system have been put forth as both a potential barrier and a potential lever for reform. This is the topic of the next section.

2.5 ASSESSMENT AND ACCOUNTABILITY

The potential for assessment to improve student learning is at the forefront of school reform. The way in which assessment can improve learning, however, takes on different meanings depending on whether this claim originates from reform based on the accountability system, such as the one advocated by No Child Left Behind, or from the reform based on visions in the Standards of the National Council of Teachers of Mathematics or the National Science Education Standards.

2.5.1 Accountability

In its objective to improve mathematics and science achievement, the U.S. Department of Education's (2002) Strategic Plan states that, through *No Child Left Behind* (NCLB), schools will be able to "use data to inform instruction" (p. 32). No strategies are given, however, on how this should be accomplished. Several recent publications (Johnson, 2002; Love, 2002; 2003; Scheurich & Skrla, 2003; Schmoker, 1996) have come to the rescue of schools with their own interpretation of how data can be used to inform instruction, although whether the data provides useful information for teachers has been questioned (Confrey & Carrejo, 2002a; 2002b; Confrey & Makar, in press). For example, NCLB advocates that teachers examine topics that students perform poorly on to beef up instruction in that area. However, it is not possible for teachers to know, based on the data they receive, whether weak performance by students on a particular topic on a test is due to poor understanding by students or a higher level of difficulty of the questions in that topic.

Moll and Gonzalez (2003) argue that one of the greatest policy threats to the ability of schools to embrace diversity is the onset of mass high-stakes testing as an educational reform strategy. They maintain that despite a lack of evidence that such testing is improving education, particularly for minorities, these tests are being developed and mandated across the nation. Furthermore, they state three unfortunate consequences of the pervasiveness of high-stakes testing in schools. One end result is

the prevalence of teaching a narrowed curriculum focused on test objectives. A second is that the overwhelming emphasis on these tests is squeezing out more useful formative forms of assessment. Finally, professional development energies are being pulled away from other issues, such as the needs of the multicultural student population, in a sense removing them from the reform agenda.

The practice of disaggregating data into ethnic and economic subgroups can ensure that schools are not neglecting students that are traditionally underserved (Scheurich & Skrla, 2001). However, Confrey argues (Confrey, in preparation; Confrey & Makar, in press) that the focus on disaggregation of data in the accountability system can also have two pernicious effects. For one, when schools only examine summary statistics in disaggregated data, it can reinforce stereotyping. For example, if teachers see the mean score or passing rate for one group, it reinforces the erroneous belief that the entire subgroup performs at that level. This stereotyping can encourage schools to rely on race-based strategies to improve scores such as pull-out programs and mandated tutoring (Kurtz, 1999) for minority students. A second, related problem with disaggregating data is that schools believe they are improving learning for traditionally underperforming subgroups (African-American, Hispanic, Economically Disadvantaged) if they see that over time they are "closing the gap" in passing rates between these subgroups and those that often perform higher. But this can be misleading as one can increase passing rates by focusing attention on those students who are closest to passing. By bringing a few students over the passing bar, for example, it appears that the entire group is improving. Focusing on this small group of students allows schools to continue to neglect both students who consistently perform well and those most at risk of failure.

2.5.2 Tensions between the two perspectives of assessment

In 1989 the National Council of Teachers of Mathematics (NCTM) put forth a vision that urged its members to consider developments in cognition and their impact on curriculum, instructional practices, and assessment strategies (National Council of Teachers of Mathematics, 1989, 1991, 1995). The National Science Education Standards (NSES) is a similar set of learning standards for Science (National Research Council, 1996) that focused on inquiry in science with the encouragement that teachers consider four areas of inquiry be assessed: precursor, planning, implementation, and closure/extension (National Research Council, 2000) with the belief that both students

and teachers would benefit from assessing students' initial ideas and following how those ideas change through the process of inquiry-based assignments and projects. Science Teaching Standard A, for example, urges teachers to "select teaching and assessment strategies that support the development of student understanding and nurture a community of science learners" (p. 22). Standard C goes further to outline five strategies for teachers in creating continuing assessments of their teaching and their students' learning: (1) to use multiple assessment methods, both formative and summative, although systematically gathering data about student understanding; (2) to analyze assessment data to guide and adjust teaching; (3) to guide students in developing self-assessments; (4) to use student data, observations, and discussion with other teachers to reflect on and improve their own classroom practice; and (5) to use student data and observations to report student achievement and opportunities to learn to students, parents, and school officials.

The National Research Council book, *Knowing What Students Know* (Pellegrino, Chudowsky, & Glaser, 2001), reports that major improvements have been made in psychometrics, but that research and applications of assessment and psychometrics have not kept up with what has been learned about cognition. Standardized tests were originally created in the mid-19th century for the purpose of checking on schools and ranking students. Little has changed over the past 150 years in these purposes. Growth in the field of cognition with a focus on differential individual ability in the early 20th century reaffirmed mental and standardized testing as ways to rank and categorize students with psychological beliefs of that era that abilities of individuals are innate and fixed across contexts and within an individual. Behaviorism in the mid-20th century further supported the ideology behind the use of skills-based standardized tests in its beliefs that knowledge is based on building up from its composite parts, learning through practice of decontextualized skills, stimulus-response methods, and extrinsic rewards. However, not much has changed in standardized tests since then to keep up with the more recent developments in cognition with constructivism and situated learning over the past 50 years that counter earlier theories of differential abilities and behaviorism. Even with the renewed interest in alternative assessments in the later half of the 20th century, far too few states have included open-ended items, portfolios, or other means to monitor the progress of its students statewide.

2.5.3 The need for developing preservice teachers' understanding of assessment

The tension between the accountability movement and the reform movement in education is of particular concern for the preparation of prospective teachers. The National Research Council (NRC) put forth twelve recommendations to the community of educators and policy-makers, at least eight of which include the cooperation of teachers and teacher educators (National Research Council, 2001a). A renewed focus on internal, rather than external assessments sends a message to teachers to reexamine the practice of a culture of "test prep", a situation often ignored by teacher educators. Colleges of education are strongly urged by the NRC to focus a major part of prospective teacher education on instruction in cognition and assessment (Recommendation 9), including a component on analyzing student assessment data:

> Typically, teacher education programs provide very little preparation in assessment (Plake and Impara, 1997). Yet teaching in ways that integrate assessment with curriculum and instruction requires a strong understanding of methods of assessment and the uses of assessment data. This does not mean that all teachers need formal training in psychometrics. However, teachers need to understand how to use *tools that can yield valid inferences* about student understanding and thinking, as well as *methods of interpreting data* derived from assessments (italics mine, p. 309).

The recommendation quoted above highlights three important skills that are needed for its realization and points to an opportunity to provide preservice teachers with these skills and experiences: a deeper understanding of classroom assessment and accountability, the statistical and technological tools skills needed to interpret student data, and an opportunity to seek inferences in the data from self-designed data investigations.

2.5.4 The need to research the effect of accountability on schools

The recent push of accountability and overwhelming use of multiple-choice tests has pressed many teachers to give up on previously tried alternative methods of assessment in favor of ones that resemble statewide tests (National Research Council, 2000). Research indicates, furthermore, that teachers are less likely to focus on inquiry if their schools are evaluated on statewide tests that assess isolated skills (National Research Council, 2000), even with outcries about the problems of standardized testing and their negative impact on instruction and disparate impact on particular student subgroups (Bernal, 2000; Confrey & Carrejo, 2002a, 2002b; Heubert & Hauser, 1999;

Klein & Remillard, 2002; Schwab, 1978b). Ball (2002) urges the mathematics education community to build a research agenda that puts more emphasis on effective assessment strategies with a clear focus on more equitable treatment of students.

The final section of this literature review examines the topic of equity, a major component of the dissertation. The research documented here forms both foundation for the equity component of the course that was used in this dissertation study and as the basis of the framework used to analyze the prospective teachers' beliefs about equity.

2.6 EQUITY

Equity issues are at the very heart of the reform in mathematics and science education over the past fifteen years. In both domains, one hears slogans such as "Everybody Counts" (National Research Council & Mathematical Sciences Education Board, 1989), "Math for All" (Allexsaht-Snider & Hart, 2001), "Algebra for Everyone" (Gamoran & Hannigan, 2000), or "Science for All Americans" (American Association for the Advancement of Science, 1989). More equitable learning opportunities are also professed to be at the heart of President Bush's *No Child Left Behind* (NCLB) legislation (U.S. Department of Education, 2001). There is a general consensus among the mathematics and science education community that "no education reform effort can succeed unless it directly addresses equity issues (Atwater, 1995; Marrett & Ziege, 1995)" (p. 11, Lynch, 2000). Yet although there is general agreement among NCTM, NSES, AAAS, and NCLB to promote equity, wording in all of these documents is too vague to help schools and teachers combat inequities in schools. Furthermore, although pockets of innovation have proven to be successful, the research community has found no single, replicable action that they can propose as a solution to the critical challenge of improving achievement for underserved communities (Campbell & Silver, 1999; Porter et al., 1990).

The context of this study is embedded in teachers' understanding of equity and fairness in testing, particularly the prospective teachers' inquiry projects discussed in Chapter 5. For this reason, the research in equity can inform the research results by situating them in the frameworks and historical perspectives of issues of equity. In this section, I will discuss some of the historical and educational issues in equity that informed this study. I will begin with a general discussion of "What is Equity?" in terms of historical and practical issues facing schools. Influences that affect student performance and their decisions to pursue high-level coursework will follow. An

overview of the subgroups who have generated the most concern is also included. Issues of equity are not well-structured problems and conflicting ideologies create tensions in working to "solve" these problems. This is a critical issue for teachers to consider, particularly if they are compelled to resolve issues of practice that may hinder equitable opportunities for their students and these conflicts are discussed next. Finally, several roadblocks to improved opportunities for students are identified in the research that allow us to begin to identify barriers that need to be broken down, and point to potential areas of improvement.

2.6.1 What is Equity?

Secada (1994) identified six categories of beliefs that teachers hold about the meaning of "equity". Kahle (1996) ordered these constructs to reflect changing beliefs in the U.S. culture over time, and linked these beliefs to specific events in history which accompanied them to demonstrate the socio-historical perspectives that may have produced these beliefs (Lynch, 2000; Thorson, 2000):

- *Post-Sputnik* (c. 1957). Equity involves maximum return on minimum investment; i.e. it advocates a concentration of scarce resources on those students who are most likely to succeed. This is aligned with Dubois' (1903) "talented tenth" idea that extra resources need to be funneled to those with the greatest promise.
- *Civil Rights* (c. 1960's). Equity is the same treatment for everyone (*equality of inputs*) so that all students have an equal chance to meet the same standards and an equal opportunity to master those standards (*equality of outcomes*). This particular belief is one that is aligned with the popular conception that equity is the same as equality.
- *Women's Movement* (c. 1970's). Equity is concern for the whole child built on the recognition that each child is an individual with unique educational, socio-emotional, and physical needs.
- *Women's Movement* (c. 1970's). Equity is a triage; i.e. investing in students most at-risk, those whose success or failure in life depends on their school experience.
- *Affirmative Action* (c. 1980's). Equity compensates for social injustice to specific groups of students who have not received fair treatment or a fair share of the resources by giving preference, when all else is equal, to underrepresented groups.

- *Diversity Movement*[1] (c. 1990's). Equity involves a safety net for individual differences, including backup programs, differentiated curricula, and other resources so that when one program does not work for a particular student, other options are available.

These six categories of beliefs about equity can be used to begin to understand some of the underlying tensions in the public discussion about what it means to create equitable schooling for students. Lynch (2000) argues that a common thread underlying these categories is three fundamental approaches to equity: equality of input, equality of output, and issues of fairness. These approaches are particularly aligned with the second and last categories of equity and articulate beliefs about the positioning of levers to create and measure change in the system.

Equality of inputs is aligned with the commonly used sports analogy of leveling the playing field. This perspective of equity is not concerned necessarily with student diversity nor does it indicate a belief that different subgroups are able to reach equal levels of achievement. Rather equity as an equality of inputs considers fairness as providing all students with the same resources, the same opportunities and supports for learning, and the same access to higher level courses. The method used to redistribute school funding in Texas, known as Robin Hood and currently under debate in the Texas legislature, is a narrowed example of this perspective of equity in its focus on providing equal funding for all students. One problem with this perspective is that it assumes that all students have the same needs and those who have fewer resources just need more of the same, when frequently the necessity is not for *more* resources, but different ones (Warren & Rosebery, 1995).

Equality of outputs, on the other hand, focuses attention on bringing minimal academic performance of student subgroups to equal levels, often measured by results on standardized tests. It does not expect all children to reach the same levels of achievement, but rather that the distribution curves of performance for different subgroups are roughly the same. Issues about closing the achievement gap often use this perspective (except that they focus on summary statistics such as means and percent passing instead of distribution curves) and this is the underlying perspective held by the *No Child Left Behind* legislation put forth by the Bush administration. Another outcome of this perspective is the idea that all children can reach minimum levels of competency

[1] This is my own designation, not in the references given.

in mathematics and science as put forth by literacy advocates and Standards documents (such as NSES, AAAS, and NCTM). One problem with this perspective is that few look for similar distributions between subgroups and instead focus on equivalent statistics, such as passing rates. The result can be that those close to passing are overly scrutinized while those at the upper and lower ends of the distribution are neglected. In addition, this view openly acknowledges a socially acceptable level of failure for some students. In Texas, for example, the schools were required to have at least 50% of each subgroup pass each test. This admits that a 50% failure rate is acceptable, making these students expendable to the system.

In the middle ground between these two is a trade-off approach that considers fairness above equality and is more aligned with Kahle's sixth perspective of equity. It assumes that schools are able to provide basic resources that ensure at least a minimal equality of inputs, but moves beyond this perspective to consider additional resources that might be needed for particular groups of students. Among these subgroups might be second-language learners, disabled students, or students with economic backgrounds that may not be able to provide rich resources or supports at home such as books, parental guidance, or additional tutoring. This perspective of equity, the ultimate goal, is also one in which consensus is not reached as to how it could be implemented (Lynch, 2000) and is contentious, complex, and likely expensive (if even possible) to carry out.

Another set of equity beliefs, based on socio-political and market forces, is put forth by Crenshaw (1988). He defines two categories of equality based on these perspectives: expansive and restrictive. Expansive beliefs focus on equality as a result. Although related to the output model put forth by Lynch (2000) above, it presumes the power exerted by the judicial system to be sufficient to eliminate racial inequalities. That is, the public's opportunity to use the law and court system ensures equity. Restrictive beliefs on the other hand, sometimes defined as a Jeffersonian view of equity, do not expect to erase inequalities but rather view equity as needing minimal input by the system. They see standards as defining outcomes that children should be able to achieve at each stage and believe that setting these standards and measuring their outcomes will exert sufficient pressure on the system to achieve the desired results; this is a belief that aligns with the Bush *No Child Left Behind* school improvement program (U.S. Department of Education, 2002). Confrey (in preparation) calls this approach the "bookends" model of accountability. Under this model, policies are put in place that

mandate that standards are set and outcomes measured but there is no effort by the system towards capacity-building. Schools and students are held accountable without systems in place to articulate or fund strategies they might need to examine, support, and improve on the curriculum, instructional practices, and policies that would enable schools to move students toward achieving the standards put in place as measured by high-stakes tests.

Secada (2000) puts forth that equity is ultimately grounded in an ethos of caring for other people. Implicit in human capacity for caring are beliefs about the nature of equality, justice and fairness, and social self-interest. These qualities are necessary to move us beyond equity as an intellectual exercise towards implementation. Several scholars have put forth definitions of equity that can be used to move the field away from more rhetoric and towards practices that exhibit Secada's ethos of caring. Lynch (2000) defines equity for science education as justice and fairness and includes sociocultural norms that support the systems that look out for those who are traditionally not included in the culture of power (Apple, 2001; Delpit, 1988). In an overview of equity in math and science reform, the Southwest Educational Development Laboratory (Powell, 1994) put forth a definition of equity that was as inclusive as possible so that "each student will be addressed as an individual, with instructional opportunities, content, and approaches that meet his or her specific needs, strengths, and interests" (p. 3). This definition appears to be closely aligned to Secada (1994) and Kahle's (1996) third category of equity. Nancy Love (p. 252, 2002), in her book *Using Data, Getting Results* asserts that:

> Equity has come to mean much more than just equal access to schools and classrooms. It implies the right of all students to reach high standards of performance. And that means their right to a rigorous curriculum, high-quality materials and equipment, a positive learning environment, and teachers who believe in their potential and are qualified to teach mathematics and science.

Although there are differences in how scholars define and use the word equity, there is general agreement among many scholars about the disservice that the polarizing nature/nurture debate to explain differential performance can cause and further aggravate inequitable opportunities provided to students. Influences discussed in the literature include Eccles' (1995, cited in Lynch, 2000) attribution theory, Valencia's (1997) deficit thinking model, Delpit's (1988) discussion of the culture of power, and Apple's (2001) claim of political influences. What is agreed is that the current methods

of schooling are not working and efforts to change focus too much on symptoms and not enough on underlying problems. This is the topic of the next section.

2.6.2 Influences

One strategy to combat inequities is to examine not just symptoms but underlying influences that may lead to differential treatment and performance of students. Lynch (2000), citing Eccles' attribution theory, argues that there are two factors that influence a student's decision whether to take advanced coursework: the student's expectation of success in the course, and the student's perception in the worth of the course either for personal satisfaction or for future careers. Intrinsic to one's expectations for success is confidence in the ability to master subject matter as well as previous experiences both in the subject itself and relative to other subject areas. For example, if students are experiencing increasing levels of confidence and encouragement in subjects such as humanities and arts, they are more likely to be "pulled" out of other areas in favor of advanced study where they have greater confidence and satisfaction, avoiding areas they perceive as risky or unrewarding (Lynch, 2000, p. 251). The other critical factor that influences students' decisions to pursue higher levels of mathematics and science is the perception of the value that those courses have for both immediate goals (e.g. personal satisfaction) and future ones (e.g. career). Much of the relative worth that students attribute to mathematics comes from socialization from parents, peers, teachers, and the media. This has been well documented in gender studies (e.g., Brush, 1991; Casey, 1996; Catsambis, 1994; Clewell & Campbell, 2002) and culture studies (e.g., Cook & Ludwig, 1998; Foley, 1997; Ladson-Billings, 1995; Ogbu, 1992; Phillips, Brooks-Gunn, Duncan, Klebanov, & Crane, 1998); it would logically extend to ethnicity, socioeconomic status, and other groups as well, although projecting findings in one area of equity into another social category has been cautioned (Secada, 1992).

2.6.3 Student subgroups of concern in equity

Historically, there has been concern in the domain of mathematics that women and non-Asian minorities (African-Americans and Hispanics) make up a disproportionately low number of those who study and make a career based on higher mathematics. In particular, research on gender has received a great deal of attention and was for a long time the focal point of study on equity in mathematics and science

education (e.g., Brush, 1991; Campbell & Storo, 1996; Casey, 1996; Catsambis, 1994; Clewell & Campbell, 2002; Pallas & Alexander, 1983; Stage & Maple, 1996). In the past decade, issues of race and culture (e.g., Ogbu, 1992; 1994; Tate, 1994; 1995a; 1995b; 2001; Valencia, 1997; Valenzuela, 1999; Weissglass, 2002) and economic status (e.g., Campbell & Silver, 1999; Orfield & DeBray, 1999) have also received increased attention in the equity literature. More recently, there has been concern that English language learners, high mobility and migrant students, and children with disabilities are also neglected in education (Heubert & Hauser, 1999). Furthermore, other student groups, although not traditionally under the lens of equity, are also of concern when it comes to representation among those who study higher levels of mathematics and science: low achievers, those with poor access to strong learning environments, and those who don't fit the science or math "nerd" stereotype (p. 10, Lynch, 2000).

A major concern about these subgroups is a perception that they are destined by their background to be unsuccessful, a perspective Valencia and his colleagues (1997) term *deficit thinking*. Deficit thinking is grounded in a belief that the reason that certain subgroups of students perform poorly is that they intrinsically possess some deficit—be it biological, cultural, economic, or family-based—that prevents them from performing at the level of their white middle-class traditional classmates, rather than seeing their poor performance as stemming from a problem with the larger educational system. As a result, teachers see these students as having personal or environmental limitations that prevent them from being successful in school. This type of thinking puts barriers on educators' perception of a student's likelihood of success and can result in these students receiving fewer opportunities to learn through lower expectations and tracking into lower level courses (Love, 2002; Tate, 1995b; Zeverbergen, 2003). These actions result in lower achievement levels which fuel the perception that students of color are less able to succeed. The claim by school administrators, for example, that "changing demographics" are at the root cause of poor school performance is a burning example of this perception.

Another key concern for subgroups traditionally underrepresented in mathematics and science is that although they may enter school with lower levels of educational experiences, there is evidence that schools may aggravate disparities in achievement by providing these students with *fewer* opportunities to acquire mathematical skills. In addition, they are most likely to experience curriculum plagued

by an emphasis on basic skills (Allexsaht-Snider & Hart, 2001; Gamoran & Hannigan, 2000; Kahle, 1996; Porter et al., 1990; Schoenfeld, 2002; Tate, 1995b).

Warren & Rosebery (1995) in their work with linguistic minority children argue that *more* of the same resources are not the same as equitable resources. The belief that children with lower levels of performance simply need more tutoring, smaller classes, and more resources implies that the current system of education is working and all that these children need is more of the same in order to become successful in the system. Rather, they put forth that learning needs to be reconceptualized to be more inclusive of diverse thinking and sense-making, particularly for children who come into the system with a different language and different culture. However, because many linguistic minority children come into the system with weaker basic skills, the pressures of the accountability system often leave urban school administrators feeling that they have no choice but to focus on improving basic skills. This means that these students frequently miss out on rich educational practices that might be available to better achieving students. In their study of teachers' sense-making in science, Warren and Rosebery found that teachers who held the belief that science was "a socially and culturally mediated process of meaning construction and criticism" (p. 303), as was evident in the scientific literature for scientists, were more likely to give their students tasks which elicited students' sense-making. On the other hand, teachers who perceived science as an accumulation of factual information discovered in a logical fashion using the "scientific method" approached their own learning about science "as a text comprehension question for which there was a 'right' answer" (p. 307) and "put greater stock in the words of an authority … than in their own sense-making" (p. 316). One can imagine how different mindsets and beliefs held by teachers about mathematics and science can go far to influence their students' confidence and value of the subject. When students don't identify with the stereotype (white, male, middle/upper class, "nerd", brilliant) that they see as being required for success in mathematics and science, they are less likely to choose these subjects for advanced study.

2.6.4 The danger of oversimplification

The notion of finding a simple solution to the equity "problem" neglects the complexity of the underlying conflicts inherent in the arguments put forth by the equity community. One might think that if one simply studies school performance data and provides lower performing students with better schools and opportunities that the equity

issues will disappear. This mindset of finding the "best" solution to a problem uncovers an underlying expectation that the issues involved are clear-cut and influenced by a single factor that can be uncovered through simple disaggregation of data and amended with straightforward (and inexpensive) policy changes. Only through examining issues in greater detail can one find evidence of their complex and conflicting factors.

Many of the pitfalls that befall those who use data to understand issues of equity can be understood by examining polar issues that accompany discussions of equity. For example, Secada (2000) argues that there is a temptation to develop stereotypes of student groups rather than focus on individuals and when studying differential performance,

> to treat demographic categories as givens. That is, once an individual is situated by her or his race, class, and the like, everything that there is to say about a person has been said. … While social groups are convenient ways of tracking how society treats people who have certain physical or other characteristics, or who are born to different levels of wealth, there is much individual variation within these groups (p. xi).

Therefore, it is critical for teachers to dig beyond superficial understandings of their students' cultures to avoid developing and solidifying stereotyped ideas of learning styles and performance. This understanding goes beyond simply identifying intellectual issues of equity, but requires probing one's own stereotypes and beliefs. It is therefore, "most important that teachers understand themselves, their beliefs and biases, and the processes by which they have absorbed their own cultures" (p. 6, Powell, 1994).

Lynch (2000) illustrates the conflict about how "best" to educate children who have traditionally performed below national norms by telling a compelling (and hypothetical) story of Elena, a bright Spanish-speaking Salvadorian girl growing up in loving, but poor family. The child is financially supported by a white man with greater means after the death of her widowed father. In the story, he struggles to find a school that would give her the best possible opportunity for a bright future, but what education would be "best"? The local neighborhood school which she will attend next year is located in an area where prostitution and drug sales are rampant, has few bilingual or certified teachers and scarce resources, and where the science curriculum consists of transmission of facts, drill-ridden assessments and worksheets. Another option would be to move her to the suburbs where public schools have better reputations. However the suburban schools are not well equipped to handle non-English speakers and often hold

lower expectations of these students who are tracked into courses where "the level of the course is inversely proportional to the number of children of color" (p. 7). A third possibility would be to send her to an affluent school with excellent resources, innovative and caring teachers, and no tracking. However this option would require Elena to ride the bus 2 hours each day, attend school with children she would struggle to socialize with, and be the poorest student at the school. "To attend this school, she would essentially have to give up her identity, or at least cordon it off from the reality of this affluent high school" (p. 8). These three options symbolize three systems of schooling that are prevalent in the United States. How can one decide which of these systems would "best" serve Elena?

The issues above put forth by Secada (2000) and Lynch (2000) highlight just two of the conflicts that arise when one engages in improving conditions for children underserved by current schooling. Polarizing and oversimplifying complex issues are dangers when one studies these issues only superficially. The complexity of issues where dichotomous yes or no decisions are encouraged (such as supporting versus disagreeing with affirmative action, choosing heterogeneous versus homogeneous grouping or tracking, or promoting neighborhood schools versus advocating desegregation through bussing) can only be understood by careful study of the contentiousness of the issues involved (Lynch, 2000). Understanding these conflicts can also help to remove roadblocks to improved conditions for children of color, English-language learners, disabled students, and others.

2.6.5 Roadblocks

Despite the research to date on equity, there is little that is understood about how one might restructure the system towards more equitable schooling (Secada, 2000). Nancy Love (2002) outlines four structural barriers and three beliefs that the literature has identified which block improvement. Structural barriers include tracking, separate classes for English language learners, Special education designation frequently given to students with behavioral or academic differences from the norm, and unequal access to gifted and advanced courses. In addition, she identifies three beliefs that also pervade schools and act as barriers in the resulting actions taken: belief in an innate ability paradigm (there are only so many 'smarts' to go around); prejudices that 'all children can learn *except* …' that come out of stereotypes of race, class, and gender; and a view of mathematics and science as a special realm for the gifted.

Tracking has long been identified as a method which aggravates differences in student achievement and increases a student's likelihood of dropping out of school (Jencks & Phillips, 1998; Orfield & Kornhaber, 2001). Research has been clear that students in the lowest tracks receive curriculum that is repetitive, focuses on basic skills, and taught by the least experienced teachers (Heubert & Hauser, 1999; Oakes, 1990). Tracking inherently embraces a tacit belief about those who can and those who can't be successful in mathematics and science. Placement in lower tracks can essentially erase any possibility that a student can later reach higher levels of study in these subjects.

There is a prevalent belief in our culture that only a few can "do" mathematics and that advanced levels of math and science and challenging content should be limited to the gifted and conducted in isolation (Lynch, 2000). Teachers' beliefs about who can "do" mathematics are largely based on their own experiences and "since most teachers have neither seen nor experienced schools delivering hard content for all students, many may be unconvinced that it is possible" (p. 17, Porter et al., 1990).

A study conducted by Seymour and Hewitt (1997) points to further problems in mathematics and science education that underlie reasons why undergraduate students leave the sciences, particularly minorities (non-Caucasians) and women. They found, as above, that some attribute leaving to lack of self-confidence, or expectation of success, or the value they give to the course both for current goals (satisfaction or enjoyment) and future careers. These elements were true for all students, but particularly for minorities. Some students felt "pulled" out of the sciences by more attractive subjects elsewhere, particularly those with exceptional talent in multiple areas. Other students felt "pushed" out of the sciences, frequently because they felt underprepared or did not identify with the competitive culture of universities that work to weed out weaker students. Above any of these reasons, what Seymour and Hewitt found to be a much greater factor for those choosing to leave the sciences was disappointment at the poor quality of teaching they found in their mathematics and science courses, and lack of engagement with the content and style of presentation by their professors. This points to the critical need to improve teaching in the math and sciences to ensure that under-represented groups are not choosing to leave for this reason. Seymour and Hewitt further report that programs designed to increase women and minority participation in the sciences have been successful at increasing entry into the sciences, but have not

been able to decrease the rate of attrition; poor instruction in mathematics and the sciences may be a major factor in this.

Beliefs underlying funding legislation are another area that can create roadblocks to improvement. Despite analytic studies that show lower opportunities for African-American students, for example, normative beliefs based on ideologies of how schools *should* function are used to sell policy instruments. Those who believe in strong government prefer mandates that force educational change although those who believe in a free-market society prefer policy instruments and standards that push the system to change on its own (Tate, 1995a). Schools, which are traditionally funded through taxes, are caught in the middle of competing beliefs about the purpose of taxation. Traditionally, taxation is used a method of inducing greater social benefit. For example, tax incentives are meant to induce behaviors that benefit the greater societal good. However politicians who embrace free-market policies support tax neutrality, a belief that the taxation system should not interfere with the decisions made by business and consumers. Furthermore, a belief that fairness in terms of taxation means that those who benefit most from a service should pay for it increases the likelihood that funding for urban schools will be shortchanged. Both of these ideologies about taxation and school funding come up short when one considers the tax base that underlies many urban and poor districts.

2.6.6 Summary of Equity

This section examined multiple definitions and perspectives of equity, as well as categories of students who are often the target of equity concerns due to these subgroups systematically under-performing on standardized tests and their under-representation in careers and advanced coursework in mathematics and the sciences. Influences underlying these elements, like lower levels of preparation or confidence and poor instruction in mathematics and science, were discussed. The problem of over-simplifying the issues in equity and seeking simplistic solutions was highlighted. This is particularly problematic when teachers are using disaggregated data from student performance to improve scores without understanding underlying conflicts. Finally, roadblocks were identified to providing equitable schooling opportunities for all children, and these implied areas of potential improvement. A critical theme in this section is that the education system, and policies created to put pressure on the system, need to change their focus from one of equity as a “problem” to one of embracing

diversity through an ethos of caring (Secada, 2000). The challenge of shifting focus of diversity into an educational asset is beautifully articulated by Ferreiro (1994, cited in and translated from the original, in Spanish, by Moll & Gonzalez, 2003) who argues that:

> It is indispensable to tool (instrumentar) the schools didactically to work with diversity. Neither diversity that is denied, nor diversity that is isolated, nor diversity that is simply tolerated. But also not diversity that is assumed as a necessary evil, or celebrated as good in and of itself. … To transform diversity that is known and recognized into pedagogical advantage: that seems to me the greatest challenge for the future (p. 699).

2.7 NEEDS IN PRESERVICE EDUCATION

The National Commission on Teaching and America's Future (NCTAF, 1996) argues that the main problem with most preservice teacher education programs is their lack of coherence, uninspiring teaching, and curriculum that lacks substance and depth. Lack of coherence and curricular depth comes from the attempts in teacher education to pack in many topics that are important for teachers to know. However, unlike the way they are needed for practice, most teacher education programs teach these skills as disconnected topics, covered only briefly (National Research Council, 2001a). Similarly, teaching in schools of education continues to be uninspiring. For example, although research has repeatedly asked for teacher educators to reform *their* teaching (Ball, 1996; Boaler, 1997; Fosnot, 1996; Polman, 2000; Simon, 1995; 2000), most teacher education programs continue to instruct through lecturing even while advocating their preservice teachers to teach through inquiry and collaborative work. Finally, the lack of substance in preservice curriculum is a result of a "once over lightly" (p. 41, NCTAF, 1996) curriculum in which subject matter is treated briefly and superficially, if at all, in methods courses.

The difficulties described above in typical teacher education courses presented an opportunity to develop and implement a preservice course that countered the complaints cited by NCTAF and furthermore, to integrate the topics in this literature review that have received calls for greater attention in teacher education: statistical reasoning, inquiry, technology, assessment, and equity. The next section will summarize each of these areas in relation to the opportunities they created for conducting this study.

2.8 SUMMARY

The literature described in this chapter highlights the need in teacher education to provide teachers with stronger content knowledge, experiences with inquiry-based learning, facility with learning technologies, deeper understanding of assessment and accountability, and proclivity towards equity. How does one provide preservice teachers with learning in all of these areas without simply adding topics to the current preservice curriculum, making worse the claims that preservice education is fraught with incoherence and lack of depth? The areas of need documented in this chapter in the literature presented an opportunity to develop a preservice course that would integrate these areas into important content for preservice teachers in an authentic context while providing them with an opportunity to learn with a reform-based approach that models the way they are being encouraged to teach. The literature provided a guide in the creation of this course.

The literature in teacher education argued that the current state of teaching in the U.S. is imbalanced towards a focus on procedural knowledge and skills and the practice of "simplifying" mathematics for students. It called for opportunities for teachers to learn content in an inquiry-based environment that modeled teaching practices that the prospective teachers were being encouraged to use when they began their own practice. In addition, it argued that the content knowledge of most teachers was insufficiently deep to carry out the reform as advocated by NCTM and NSES. These issues, together with the model of the National Writing Project, set forth an assumption that prospective teachers needed an inquiry-based environment to learn content, in this case statistics, by *doing* statistics rather than focusing on teaching them how to teach statistics.

The statistics education literature pointed to the parallels that sometimes exist between the development of children's reasoning and that of novice adults. An emphasis on distribution and variation was shown to be a critical area in need of development in the school curriculum as well as a focus on conducting statistical inquiry rather than statistical rules and procedures. This provided evidence that statistics was an important area of content for the preservice teachers to develop. Attention to the difficulties that adults have been shown to have in breaking a deterministic mindset in applied problems emphasized the importance of immersing the preservice teachers in statistical learning within an authentic context, so that important statistical concepts were linked to their uses as tools for inquiry. The recent attention and success of

supporting children's initial informal language and concepts and the potential dangers of building rote knowledge through simply teaching standard graph forms and statistical measures opened up the possibility of paralleling a more informal alternative for preservice teachers in their understanding of variation and distribution. In addition, the challenges documented in children's development of data-based inquiry re-emphasized to the importance of providing teachers with this experience. In addition, the pilot study, conducted with practicing teachers, and which brought together these important areas of need in statistical reasoning, provided a testing ground for much of the material used in this study.

The inquiry literature went beyond simply reiterating the importance for teachers to experience inquiry-based learning. It also set up a framework and set of critical categories to attend to as teachers conducted their inquiries: formulating questions, making and investigating mathematical conjectures, developing mathematical arguments, evaluating predictions based on data, building mathematical knowledge through problem-solving, and communicating mathematical thinking clearly to others. In addition, it pointed to the importance and difficulty of developing a mindset of inquiry and the importance of using authentic, ill-structured problems to provide experiences with uncertainty and complexity in inquiry to assist with the development of this mindset. In addition, the literature emphasized the potential for authenticity and uncertainty to increase the prospective teachers' interest and cognitive motivation to conduct the inquiry.

In addition to its potential for supporting and motivating reform-based practices, the technology literature provided insight into the dangers of teaching teachers about technology without giving them enough experience to develop facility with it. The overuse of technology, especially by teachers who are less comfortable with computers, for diagnostic purposes and skills practice, particularly for students in poor communities, re-emphasized that facility with powerful, dynamic learning technology was important for the sake of equity of the preservice teachers' future students.

Problems documented in the literature on assessment and accountability was one of the major motivators for using this as a context for the teachers' inquiry projects, particularly in light of its importance for equity. The equity literature provided both insight into areas of importance to discuss with the prospective teachers and as insight into the kinds of beliefs about equity that the prospective teachers might have. The

previous work by SYRCE in uncovering many of the downfalls in schools' uses of disaggregated data provided critical categories of investigations and junctures for discussion that were used in the development of the course.

Taken as a whole, the literature provided important insight into how the course was organized into a coherent set of integrated themes. It also documents how the course and areas of study chosen in this dissertation are grounded in the literature. The next chapter will describe the course in which the dissertation study took place as well as set up the theoretical assumptions, design of the study, and method of analysis used.

Chapter 3: Theoretical Basis, Design, and Methodology

This chapter outlines the theoretical basis, design, and method of analysis of preservice teachers' understanding of variation and distribution in articulating issues of equity and fairness in the accountability system. The purpose here is to examine where these two general topics – statistics and equity – overlap and interact to provide mutual support for the development of each. This study is meant to focus on this interaction rather than provide a systematic study of teachers' understanding of either topic independently. This is a new area of research, but one that builds on previous studies in a variety of areas, as was discussed in the previous chapter.

The study is informed by a set of assumptions, values, and beliefs held by the researcher about teaching and learning, epistemology, equity, and research. This chapter begins by articulating the theoretical bases of the study. Secondly, it sets out the design and preparation of the study, based on a methodology identified as *design research* (Brown, 1992; Cobb et al., 2003; Confrey, 2002a; Edelson, 2002). Next, the details of the conduct of the study—subjects, setting of the study, and data generation—will be described. Finally, the method of analysis, based on Grounded Theory (Strauss & Corbin, 1998), will be summarized.

3.1 THEORETICAL FRAMEWORK

The assumptions, values, and beliefs of the researcher not only motivated the choice of the study, but also influenced its design and methodology, including the intermediate decisions made, conjectures and refinements, choice of data collected, method of analysis, and organization of results. This section will disclose my assumptions, values and beliefs in the following areas: epistemology; teaching and learning, particularly as it relates to statistics; values and beliefs about equity; and conceptions about research.

3.1.1 Beliefs about Epistemology

Epistemology is the study of the nature of knowledge and justification (p. 97, Howe, 2003 citing Moser, 1995). I agree with Confrey's description of the constructivist paradigm (Confrey & LaChance, 2000) that "mathematics is viewed as a result of personal constructions made through one's actions and reflections on those

actions." (p. 238). This perspective holds that knowledge is not passively received, but rather actively constructed by the individual. Although knowledge is a human construction, it is not constructed by the individual alone, but also through mediation with social norms. Individuals may hold competing beliefs that are logically contradictory. I reject the positivist paradigm that claims that there exists a single truth with its own independent existence, one which can be "discovered" and externally observed, and that knowledge is gained through transmission. Knowledge, however, is not completely subjective and relativistic to the individual, but negotiated by social norms of practice.

I further hold that the world is complex. Therefore, a diversity of perspectives, through democratic debate and transparent discussion, scrutiny, and argument, can move the multiple individual viewpoints and understandings about the world toward a more richly shared set of goals that respect diversity. Furthermore, a diversity of ideas ensures that the goals reflect not only the community as a whole, but also protects and gives voice to those frequently silenced by the majority.

3.1.2 Beliefs about Teaching and Learning

In the majority of American mathematics classrooms, the emphasis is on learning formal mathematical terminology and procedures (Stigler & Hiebert, 1999). Because students are frequently taught procedures without first developing their own intuition and reasoning about the concepts underlying the procedures, the teaching of formal rules has in many cases lead to students suspending their own sense-making, thus inhibiting their understanding (Flyvbjerg, 2001; Schoenfeld, 1991). I believe that formal mathematical procedures, terminology, and symbolism *are* critical to developing mathematical understanding in that they can provide more efficient paths to problem-solving, focus attention on particular aspects of a problem, and open new levels of understanding of the concepts represented by the terms or symbols. However, the emphasis must be on building sense-making, not simply using procedures or terms when their underlying purpose is not understood. Developing that understanding oftentimes requires leaning on non-standard terminology and approaches to problem solving before one can use standard ones (Lemke, 1990). This belief is evident in the development of the preservice course in which the subjects in the study were enrolled; that is, sense-making was the larger goal, while standard terminology and procedures were secondary.

Constructivism further implies that teachers need to provide students with authentic tasks that allow them to construct their understanding. This does not mean that students should be left to "discover" for themselves the mathematics that has taken society thousands of years to develop. It is desirable that students develop the kinds of mathematical and scientific practices that their mathematical and scientific communities have agreed upon. It is also desirable for them to understand, as Kuhn (1961/1996) has articulated so well, that the paradigms in which scientists (and mathematicians) work are not "truth", but are subject to change.

3.1.3 Beliefs about Equity

I acknowledge that people hold different beliefs about equity and that these beliefs are largely influenced by the experiences and culture that people engage in. I also acknowledge that understanding of equity, like understanding of mathematics, needs to be constructed by an individual, in the company of others. As a teacher, I cannot "tell" a student what to believe, but I can nurture a classroom culture and organize opportunities where students can reflect on and openly discuss beliefs. I do believe that some understandings about equity are "better" than others. The equity framework developed by Secada (1994) and ordered to correspond to historical development by Kahle (1996) includes six common beliefs about the meaning of equity. I agree with Kahle's ordering of these beliefs; the historical ordering that she developed corresponds to my own preferential ordering. For example, I believe that equity as a safety net—so that if one approach does not support a student's understanding, other resources are in place to support that student—is a "more" equitable approach than the biggest-bang-for-the-buck belief that resources should be concentrated on those students most likely to make the largest contribution to society.

3.1.4 Beliefs about Research

Different methodologies and research designs serve different purposes in education. Although the standards of research in education are the same as those in other areas of study, the complexity of educational settings create additional challenges, particularly for studies that aim to aid educational practice (National Research Council, 2002). Although it is acknowledged that situational elements of the setting play a larger role in educational research, there is still much to be learned in a context-laden study. The researcher as well as the reader of the study can speculate, based on the evidence

and description of the study, on aspects of the context that may have played an important role in the outcome of the results. These elements can inform other research studies to be sensitized to these contextual factors and if they emerge in multiple studies, the argument is strengthened that they should be further researched in broader settings. For example, several research studies have implied that the context of learning plays an important role in the depth and breadth of learning that occurred, and have also documented ways in which context might be an obstacle (Bransford, Brown, & Cocking, 2000). No single type of study can provide the research community with all that it needs to understand and improve practice, however, a diversity of approaches can work together to inform one another.

Under the assumption that the world is a complex place for which there is not one definitive "truth", research takes on a different meaning for me than it might for those who hold exclusively to the belief that the purpose of research is to make definitive claims about the way that the world "works". I reject, for example, the definition of "scientifically-based research" that is made by the current Bush administration (U.S. Department of Education, 2003), one which excludes from its definition educational studies that work to develop theory and new approaches to learning as opposed to ones that evaluate pre-existing theory. By allowing for multiple research designs that serve multiple purposes, several benefits accrue. First, the very nature of multiple approaches to research provides the community with multiple levels of understanding and diversity of approaches. Second, common elements and themes can be observed in one setting which allow researchers to be sensitized to their existence in other settings. Finally, by articulating the purpose and assumptions that undergird a research design, other researchers can determine which elements are convincing and applicable to their own design. Furthermore, practitioners can pick up on those elements that fit with their own beliefs about classroom practice, making it useful to a broader audience.

I agree with Flyvbjerg (2001) in that social science research should be built on values that encourage the researcher to include, as part of the research, wisdom developed through deep understanding of context. Flyvbjerg lays out three kinds of understanding: *episteme, techne,* and *phronesis*. Episteme is reflected in theoretical knowledge, which is by nature context-independent. By itself, if research focuses on theory, it constrains the possibility of its use in practical settings. Therefore, sole

reliance on context-independent knowledge can impede progress in understanding and improvement of practice. Techne, or practical "know-how", focuses on the outcome or end-product of knowledge application. It is by nature context-dependent and can be equally limiting if relied on exclusively. Personal experience and contextual knowledge, on their own, lack the overarching understanding needed to prioritize and plan for understanding in multiple contexts. According to Flyvbjerg, phronesis, the third type of knowledge, is rooted in values and ethics. It takes into consideration theory, context, and practical experience, but is also deeply rooted in the desire to improve society. The study developed here rests on the assumption that in order to understand the connection between statistical understanding of variation and distribution and equity and fairness in the accountability system, one must pay particular attention to the context of the setting insofar as it informs understanding that moves society towards improved practice. In this case, practice is assumed to be the classroom practices of teaching and learning, but also can be extended to inform understanding of the greater society in which the classroom is situated. These assumptions mean that any conjectures and theories that are developed are humble, subject to changing conditions that emerge even in the course of the study. The purpose of this study is not to confirm a pre-existing theory, but to examine closely the connection between preservice teachers' proclivity towards equity issues and their understanding of variation and distribution, and the environment designed to promote it.

Educational research by its nature is value-laden and subjective, with the aim of improving practice. Without research based on values and ideals, "we fall into a hopeless kind of relativism … [that] evaluates [empirical research methodologies] not by *a priori* epistemological standards, but by the epistemological standard of their fruitfulness in *use*" (p. 11, Howe, 2003). If educational research is to rely on a phronesis that honors democratic ideals, it must also submit to three principles of democratic research, as described by Howe—inclusion, dialogue, and deliberation. Inclusion implies that the researcher pays attention that the sample of participants is representative of a diverse set of views and that these diverse views have the right to be included if the results are to have bearing on social practice. The principle of dialogue requires the researcher create a setting where the views reported by participants are genuine, ensuring their authenticity. The third principle put forth by Howe is that of deliberation. This principle looks to foster equality in dialogue and pay attention to the

conditions in which it can emerge. Views that are expressed must be given the opportunity to be clarified, for example, through probing, and the self-understanding of subjects in the study is valued by subjecting their views to rational scrutiny. Bias is controlled under these principles through the inclusion of all groups to whom generalizations are intended, and include the genuine voices of each group. This is in contrast to much of cognitive research that aims to control the setting to exclude contextual factors. For example, the purpose of the clinical interview is to subject each participant to an identical setting and context to ensure that contextual factors do not confound results. While I agree that some control of setting provides additional insight through the opportunity to compare responses, a more clinical approach may not work hard enough to seek to understand the complexity of experiences brought into the interview by the individual. I hold that a balance between a clinical setting and a focus on individual cases can also provide additional insight by gaining understanding of commonalities and differences in responses under similar circumstances, and at the same time allowing for individual experiences to add to the explanatory power of the results, resulting in a richer data set.

3.2 PILOT STUDY

The pilot study, conducted in the spring and summer of 2001, allowed me to test several of the activities and assessment items (see Chapter 2 for description of the pilot study). Although the pilot study involved practicing teachers rather than preservice teachers, it was presumed that the similarities between the two groups would be sufficient for the purpose of testing and refining the study. Detailed description and findings of the pilot study, which were included with the dissertation proposal, is available in Appendix A.

3.3 FATHOM

The software chosen for the project, *Fathom*™ (Finzer, 2001), is unique in its application as a teaching and inquiry tool. Whereas most statistical software tends to be like a "black box" (data in, answers out), or designed for very specific kinds of tasks, Fathom can be used to investigate a broad range of tasks at both an elementary and intermediate level. In addition, many schools in the district had already purchased Fathom (although it was not yet widely used) and this would presumably increase the

likelihood of the prospective teachers using Fathom with their own students when they began student teaching and later in their careers.

Software instruction was provided on graphs and statistical summaries, importing data, least squares regression, relational graphing, sampling, and simulations. At an informal level, Fathom allows its users to quickly test simple conjectures within the first few minutes of use with the software. Its ability to easily "drag and drop" variables onto graphs and to be able to link relationships simultaneously in several graphs made it easy to begin using inferential language with the prospective teachers from the very beginning. For example, in the pair of dot plots below (Figure 3.1, adapted from Confrey & Makar, in press), the Math TLI scores of a group of 7th graders at a local middle school are shown in the top distribution. The Math TLI scores of these same students the year before are represented in the distribution below. When the students near passing (TLI = 70) in grade 7 are selected (shown in black), one can observe and begin to suggest generalizations about the scores of these students in consecutive years. Usually, one assumes that the students' scores in grade 7 will be similar to their scores the year before and that students with similar scores one year will have fairly similar scores the following year. This representation counters that stereotype.

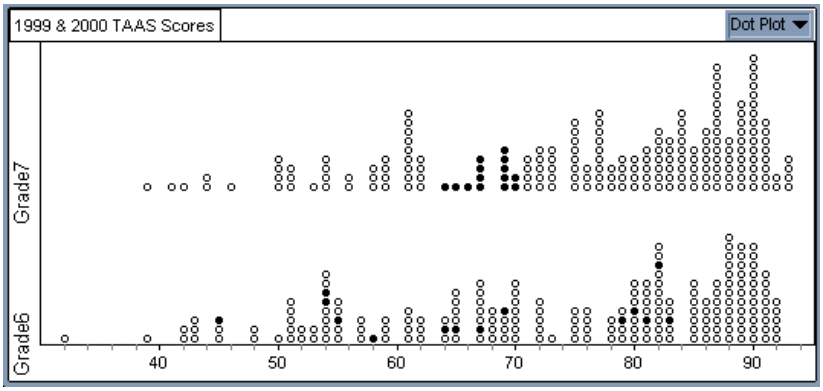


Figure 3.1: Math TAAS scores of students in seventh grade at a school (upper) are compared with the same students' scores the year before. Students close to passing in grade 7 are highlighted to investigate the previous performance of these 14 students. (Data source: school administration, 2001)

Advanced levels of the software can be used to set up simulations to conduct formal or informal hypothesis tests and estimate parameters through more visual means than is possible with the output of a significance test or calculation of a confidence interval. For example, in comparing two groups, even when there are no systematic differences between the distributions of the variable being compared, the means of these two groups are likely not identical. How does one decide if the difference is likely due to natural variation or if there is evidence of systematic differences between the groups? A standard method of doing this in statistics is with a significance test. The underlying concepts and mechanisms behind this procedure, however, are hidden from the user and a statistics student may not understand the assumptions and reasoning behind the test (Abelson, 1995; Reichardt & Gollob, 1997). In Fathom, one can build a simulation that models the null hypothesis that there is no difference between these groups by "scrambling" the group characteristic in the two groups and comparing the original difference in groups with the difference in the scrambled one (repeated many times) to get a sense of how unusual the observed difference would be if natural variation were at play. For example, the observed mean difference between math SAT scores for a sample of about 500 males and 500 females was about 43 points. Under the null hypothesis that there is no difference in scores, the range of differences in scores between these groups would most likely fall between –10 and +10 (Figure 3.2). Therefore, the likelihood of a difference of 43 points occurring just be chance would be extremely rare.

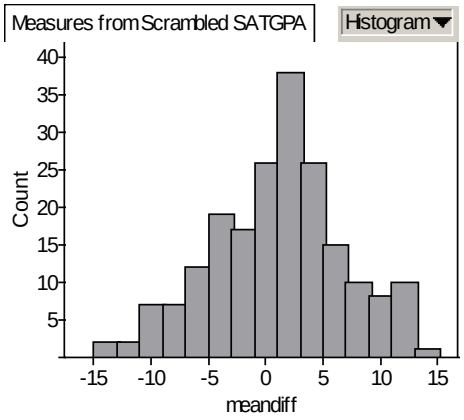


Figure 3.2: A simulation showing a likely distribution of differences in SAT math scores between 200 samples of 500 males and 500 females if there were no difference in the population.

The affordances provided by the software were critical to engendering the intuitive, conceptual development of the prospective teachers' understanding of variation and distribution. The ease of creating graphs to envision relationships within and between variables or subgroups also afforded an opportunity to shift emphasis from numerical summary statistics towards interpretation of more visual representations. Because the emphasis in the preservice course was meant to focus on these more interpretive aspects of the data, the choice of Fathom was made for the project.

3.4 DESIGN OF THE STUDY (DESIGN RESEARCH)

In light of the beliefs about epistemology, learning, equity, and research articulated above, I have chosen a design for the study that highlights development of theory, iterative process of refinement, applicability, and democratic values. This type of design implies a need to engineer an innovative learning environment that promotes understanding both of the learners being studied, and the learning of the researcher conducting and revising the study. In addition, if the study is meant to inform other learning environments, the context must be a practical and authentic learning environment where the researcher simultaneously studies and works to improve the learning of subjects. Several educational researchers have relied on a kind of research design that holds these ideals, called *design research* (Cobb et al., 2003; Confrey, 2002a; Edelson, 2002) or *design experiments* (Brown, 1992). The use of the word "design" in the title is meant to emphasize the focus on the design, or engineering, of the learning environment.

While a design research study may focus on a defined set of particular aspects of the setting, it does not ignore the larger system in which the setting is situated as a systemic whole. It acknowledges that changes to part of the system can have significant impact on other aspects of the system. Ideally, the research study, through inclusion of a research team focusing on multiple levels and aspects of the system, can work to understand not just the focus of each part of the study, but also the greater system. No amount of resources, however, can permit all aspects of the system to be included. Therefore, aspects under study are by nature limited by available resources and choices made by the researcher, as well as those restrained by the setting. Design research, by design, includes an intervention. The goal, according to Brown (1992), is "to work toward a theoretical model of learning and instruction rooted in a firm empirical base … that not only work by recognizable standards but are also based on theoretical

descriptions that delineate why they work, and thus render them reliable and repeatable" (p. 143). Because the intervention engineered to study the learning environment often includes designing innovative curricula, norms of classroom practice, and assessments and recognizing that one aspect of the system can influence others, yielding results that can be confounded in ways that would frustrate many empirical researchers. Although the process of integrating theoretical and empirical bases of the study can create methodological problems, results can be considered more valid in that they occur in the complex and constantly changing arena of the classroom.

Cobb et al. (2003) identify five features that are common to design studies: (1) a overarching focus on the development of theories that support learning; (2) the implementation of an innovative intervention that is designed to seek factors that contribute to targeted forms of educational improvement; (3) a simultaneous prospective and reflective process that begins with a hypothesized learning process and support system that is continuously scrutinized, and then undergoes reflection, capitalizing on contingencies that emerge as the design unfolds; (4) a dependence on iterative design that refutes, revises, or refines the conjectures being tested, the planned learning environment, and parallel measures created; and finally, design studies include (5) the acknowledgement that theories developed by this process are humble, intermediate, and specific to the design. The great advantage of this process is that,

> in contrast to most research methodologies, the theoretical products of design experiments have the potential for rapid pay-off because they are filtered in advance for their instrumental effect. They also speak directly to the types of problems that practitioners address in the course of their work (p. 11).

3.4.1 Preparation of the Study

The preparation of the design experiment for the dissertation study included all five factors articulated by Cobb, Confrey, diSessa, Lehrer, and Schauble (2003):

1. *Overarching theory to support learning*. The point of the study was to research the interaction between prospective mathematics and science teachers' use of statistical concepts of variation and distribution and their understanding of issues of equity and fairness in accountability through the process of data-based, technology-intensive analysis and inquiry. The initial conjecture was that teachers' conceptual understanding of variation and distribution would influence

their understanding and proclivity towards equity. This theory was revised during the experiment.

2. *Intervention to support educational improvement*. As laid out in the literature review, there is a great need for teachers and schools to make better decisions about student learning within the accountability system. The challenge was to support and test the theory while at the same time develop an intervention which would support and improve teachers' use of accountability data. The opportunity and support for experimentation with the preservice course was likewise critical to ensuring the authenticity of the study setting. Although the purpose of the research was not to evaluate the intervention directly, it was expected that the implementation of the study would have an impact on the theory under development. The intervention designed to deepen teachers' understanding and use of data was created in the form of an innovative preservice course on assessment, data analysis, equity and inquiry. These topics were taught both sequentially and concurrently. As each topic was introduced, it was built on and integrated with the previous ones while simultaneously foreshadowing and opening the space for initial conceptual development of the next topic.

   The course began with an introduction to assessment (from classroom formative assessment to standardized testing), while simultaneously exemplifying assessment with data use, pointing out issues of fairness, and modeling a mindset of inquiry. The preservice teachers were then introduced to statistical concepts, through analyses of data, to deepen and expand their understanding of assessment—its multiple purposes, interpretation of results, and limitations. Data analyses were presented by the instructors and discussed in class in initial stages, and continually modeled throughout the course. The preservice teachers then embarked on exploring and conducting their own analyses. These began with structured assignments which became increasingly open-ended over the course to prepare them to design their own inquiry.

   Opportunities to discuss equity, even though it had not been introduced formally, were capitalized upon. Equity was introduced formally in the second half of the course after the preservice teachers completed a required 3-day teaching experience in a local urban school (as part of the preservice program). Readings in equity were usually set in the context of accountability, and

designed to encourage the prospective teachers to broaden, reflect on, deepen, and articulate their personal beliefs about equity publicly. During this unit, set in the context of assessment and accountability, the prospective teachers integrated and improved their data analysis capabilities with their emerging understandings of equity.

3. *Prospective/Reflective Process*. The initial conjecture and learning activities designed to support it were planned at the beginning of the course, but then revised throughout the semester as the study unfolded. The instructors met after each class to debrief, reflect, and discuss ideas to strengthen or revise the plan for the next class.
4. *Iteration*. Several levels of iteration took place during the course of the study. On an implementation level, the prospective/reflective process described above was iterated and greatly influenced details of the execution of the course. Because the implementation of the course was new, instructors had to revise daily plans as they developed a better understanding of the prospective teachers' thinking about each topic. On a research level, conjectures were revised to capitalize on emerging opportunities and adjust due to limitations observed and contingencies that arose. These revisions also had an impact on the course design and implementation. Additionally, intended research measures had to be adjusted or new ones created.
5. *Putting the theory to work*. Critical to the theories being developed were the simultaneous implementation of the course and the research study. The course was developed not just for the dissertation research, but also served as a pilot for a course that could be used for preservice teachers or even graduate students wanting to better understand testing and accountability. Although the study is meant to inform the research community, the course can be (and was) tailored to a particular setting and implemented almost immediately at another university (Confrey, Makar, & Kazak, 2004). In addition, the application of these results to practice took the form of feedback of the UTeach preservice teacher education program with suggestions for improvement. Furthermore, the results were reported to researchers studying statistical reasoning (Makar & Confrey, 2003; under review).

Research Questions

The pilot study was designed to promote and study teachers' statistical reasoning, ability to conduct a reasoned investigation of student assessment data, and deepen their understanding of the context, opportunities, and limitations of high-stakes testing. What was observed in the pilot study, and serves as a major focus of the dissertation study, is how some of the teachers used their deeper understanding of statistics, particularly variation and distribution, to argue for more equitable treatment of students. With a greater focus on this interaction, it was conjectured that a better understanding could be developed of elements that promote and discourage this connection. With that in mind, the central research question for this study was:

In a preservice course created to support learning about assessment, technology-driven data analysis, equity, and inquiry, how do prospective teachers use the concepts of variation and distribution to support their understanding of issues of equity and fairness in testing?

In order to unpack the central research question and better understand elements that might be central to the interaction, four sub-questions were created:

1. What level and types of understanding of the concepts of distribution and variation were learned? How did the teachers express this understanding in practice?
2. How was the technology used in relation to the students' inquiries? What behaviors did the prospective teachers exhibit in using the technology in a semi-structured investigation?
3. What can be said about preservice teachers' understanding of equity from their structured and ill-structured inquiry activities?
4. What is the interplay between the preservice teachers' statistical reasoning and the depth and breadth of self-designed inquiry into complex, ill-structured problems?

The first sub-question focused purely on the prospective teachers' statistical understanding, one major element of the interaction. In order for teachers to *use* statistics to understand equity it was necessary to understand the intricacies of their understanding of the content. The second sub-question probed the potential influence of the technology on the interaction. Although it was not a major element of the research question, it was clear that the investigations the teachers would conduct would not be

possible without their facility with the technology. It was therefore important to understand the affordances and constraints of its use. Equity, the focus of the third research sub-question, was the other major element of the interaction under study and it was critical to gain insight into the teachers' beliefs about equity in order to understand how they would play out in the interaction. Finally, the use of inquiry in a complex, ill-structured problem was conjectured to play a critical role in the way teachers used statistics as evidence to support their investigations of equity. This was the focus of the fourth sub-question.

Because the research question itself was so complex, it was necessary to break it down into the smaller sub-questions. However the diversity of these sub-questions, increased the danger that the additional breadth of the study would compromise the depth of understanding in any one area. The decision to focus on breadth was determined to be critical for a study in such a new area of research in order to ensure that major factors contributing to the interaction under study were considered. In addition, the breadth of the study would enable researchers with specializations in each of these areas to use this study to probe their own area of research further and take advantage of their expertise in a narrower focus. Furthermore, the opportunity for this study to inform practice would have been compromised by a narrower and deeper focus, as all four elements are critical to the interaction being studied and their implementation into practice. Some additional data was collected in order to allow the researcher to probe into related areas in more depth than possible here for later research, particularly in statistical reasoning and beliefs about data-based inquiry.

3.4.2 Conduct of the Study

Subjects

The subjects of the study were preservice mathematics and science teachers enrolled in Classroom Interactions, the second of three required education courses in the UTeach program for prospective secondary teachers at the University of Texas at Austin. The course began with 22 students and four students were lost during the semester due to attrition; three students left the university altogether, while the fourth transferred after a few weeks due to a scheduling conflict. One additional student finished the course, but did not turn in her final project. The sample was selected as a convenience sample; therefore, because the sample was not chosen randomly, it cannot

be generalized to the larger population, either within the UTeach program or to the larger population of preservice math and science teachers in the United States. Consultation with a mentor teacher in the program who knew all of the students indicated that those enrolled in the study section were representative of the students in the UTeach program. It should be noted that UTeach is a collaboration between the College of Education and the College of Natural Sciences and all students in the program are required to major in mathematics, science, or computer science. Therefore, it is likely that these students have a higher level of content knowledge than the population of secondary math and science preservice teachers in the United States.

Although the sample for the study was a convenience sample, there is likely no systematic selection bias as (1) students who enrolled in the course did not know that the class would be any different than the regular sections and (2) the only reason to select this class over another is scheduling convenience, which probably does not result in a systematic bias. In addition, the composition of the students in the course were such that 39% were minority students (African-American or Hispanic), slightly over the UTeach average of 26% (LaTurner, 2003) which allowed a representative "voice" for minority students. The gender breakdown in the preservice course was heavily female with 83% of the course women compared to a 62% average for UTeach.

The advantages of the change in population between the pilot study and the dissertation study, from practicing teachers to preservice teachers, are that: (1) secondary preservice teachers majoring in math and science would likely have a stronger overall content knowledge in mathematics than typical practicing teachers, and potentially be able to gain greater depth of statistical understanding by the end of the study; (2) preservice teachers, mostly undergraduate students, are likely less resistant to learning new content than practicing teachers who have been out of school for a while and may see an attempt to teach them content as a criticism of their content knowledge; (3) because they are enrolled in a university course, they would be a more stable population than the pilot study which suffered from a high rate of attrition; (4) it is very difficult to gain extended access to practicing teachers; and (5) it was conjectured that promoting understanding of preservice teachers might have a greater long-term impact on students given that preservice teachers, who are university students, may be more flexible and open to new ideas than practicing teachers. The major drawback with using

preservice teachers was that they might reject the need to understand issues of testing as they had not yet experienced the pressures of the system to improve student learning.

TAAS

Until 2002, the Texas Assessment of Academic Skills (TAAS) was the state's multiple-choice test of the statewide curriculum in reading and mathematics in grades 3-8 and 10. In addition, TAAS writing tests were administered in grades 4, 8, and 10. Students were required to pass the exit-level test (grade 10) in order to graduate. Test results were shared with teachers and schools in the form of summary data and hardcopies of individual student performance. The data are also disaggregated and schools and teachers held accountable for their students' performances in four particular subgroups: White, Hispanic, African-American, Economically Disadvantaged. If any of these subgroups, having at least 30 students and comprising at least 10% of the students being tested, fell below 50% passing, the school earned a "low-performing" rating and was subject to state sanctions. Schools could also be rated low-performing based on their dropout rate by each of these subgroups. Schools with passing rates exceeding and dropout rates below the state requirement could earn higher ratings: Acceptable, Recognized, and Exemplary. The student-level criteria for passing TAAS was defined as having a Texas Learning Index (TLI) score of at least 70. The TLI was a scaled score, converted from the student's raw score, that is determined each year by the state. Although the overall passing rate of TAAS had increased over the years (from 61% in 1994 to 93% in 2002 on the mathematics test), the percentage of questions required to pass fell over the same period, ranging from 70% (1994) to 48% (2002) of 60 mathematics questions needed to be answered correctly. This puts the claim up for debate that the accountability system is improving student learning in Texas (Confrey & Makar, in press).

Setting

The setting of the study was a one-semester course for prospective teachers on classroom instruction and assessment called *Classroom Interactions* developed by Jere Confrey. This is the second of three required courses for students in the UTeach program, a secondary mathematics and science teacher preparation program run by a collaboration between the College of Education and College of Natural Sciences at the University of Texas at Austin. The study was conducted in a special section of the

Classroom Interactions course with a particular emphasis on assessment and equity through data-based inquiry using Fathom (Finzer, 2001) and was co-taught and co-designed by Prof. Jere Confrey and myself (Confrey, Makar, and Kazak, 2004). The course ran from January through May 2003. There were three sections of the course available for preservice teachers to choose from and the subjects of the course were those students who signed up for one particular section assigned to Dr. Confrey. They did not know when they enrolled in the section used for this research that the course would differ from the other sections offered. The first day of class the students who enrolled in Dr. Confrey's section were informed about the research study, given the chance to switch sections, and asked to sign a waiver.

The course was modified from the content of a customary *Classroom Interactions* course. The major themes of the original course—pedagogy, assessment, and equity—were kept in place to ensure that the preservice teachers did not miss the major ideas of the intended syllabus. For example, a major portion of the original course was dedicated to the planning, carrying out, and reflection of a three-day teaching experience at a local high school. This element of the course was also preserved, even though it was not expected to benefit the study.

The design of revised course used for this study had two purposes – to research the interaction of teachers' understanding of equity and variation, and to create a classroom environment to support and promote this understanding. When at odds, the learning goals took priority over the research goals, however, the research goals were designed to minimize interference with the regular workings of the course.

Most of the instruction in pedagogy for these teachers included a strong component advocating for inquiry-based learning with innovative technology. A major goal of the course was that the teachers be able to experience statistical inquiry as learners before they would enter their own classrooms and teach inquiry-based instruction. The major themes of the course—assessment, classroom instruction, equity, and inquiry—were interwoven throughout the semester, but each was emphasized during a four-week unit in the course. An overview of each unit is given in the syllabus (Appendix B).

The capstone of the course was an inquiry project into an issue of equity or accountability. The students presented their findings to the class on one of the final two days of the course. This project comprised a major portion (40%) of the course grade

and served to synthesize the readings, teaching and learning experiences, resources, and discussions from the course as well as draw on their experiences during the course, and specific interests.

Statistical Content

The scope and sequence of the statistical topics were not what would be considered conventional. Instead, the focus of the course was shifted from a fixed body of content to statistics as a tool for analyzing data to gain insight into particular issues. Most often, data were examined through graphical representations (dot plots, histograms, box plots, and scatter plots) and much of the discussion centered on a particular representation highlighting issues in equity and accountability. Throughout the course, distribution was a key component of discussions and teachers were encouraged to interpret distributions not as mathematical objects, but as tools to think and debate with. For example, the teachers discussed what might lead to a distribution being skewed (perhaps because of a ceiling effect of an easy test), as in the dot plot of state test scores from sixth graders at an urban middle school shown below (Figure 3.3, below left). The software allowed for cases selected in one graph (right) to highlight in other representations (left), for example in this pair of graphs which show the socio-economic status of students highlighted to show their test scores ("1" means they've applied for the economic assistance through the national free lunch program). The display shows prospective teachers that academic performance on MATLI (the math scaled score for the Texas Assessment of Academic Skills) is widely distributed for economically disadvantaged students.

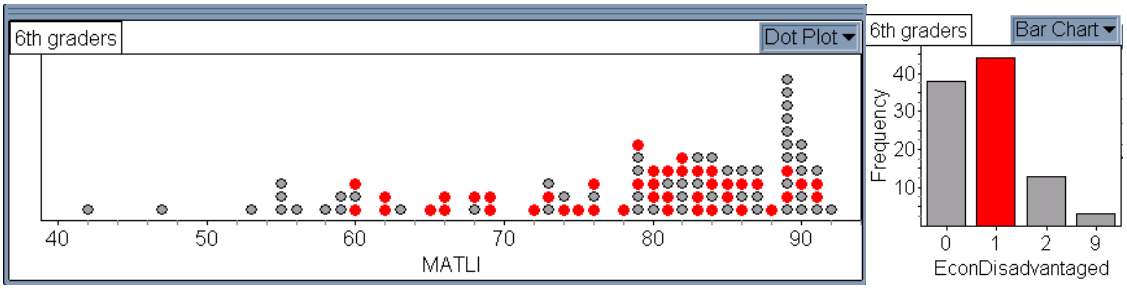


Figure 3.3: A plot of student scores on TAAS-math with the economically disadvantaged students highlighted.

Additional content discussed was distribution shape, measures of central tendency (mean and median), standard deviation and quartiles, association and correlation of bivariate data in scatterplots, basic linear regression and residual analysis, and informal concepts of inferential reasoning (null hypothesis, sampling distributions). The purpose was not to be comprehensive in treatment of topics, but rather use the statistical tools to illustrate concepts in data. The content was not formally *taught* to teachers (except during their preparation for teaching a unit on regression at a local school), but came out of class discussions highlighting a particular topic in equity or assessment. For example, sampling distributions were used to illustrate the probability that a small subgroup of students might fall below 50% passing, causing the school to undergo sanctions, even if their predicted scores were 54% passing (Figure 3.4) based on past performance trajectories. In this context, teachers also had the opportunity to investigate how this probability was dependent on the size of the subgroup.

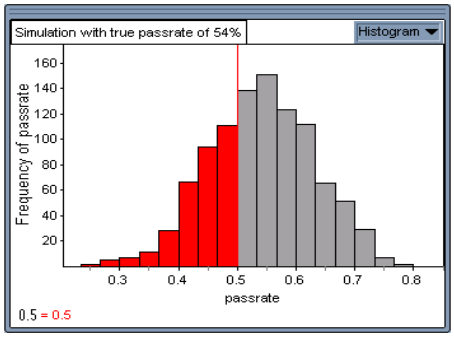


Figure 3.4: A simulation showing the probability of a small student subgroup falling below state requirements even if their "true" passing rate is above it.

Investigations

Two types of data investigations took place during the course, those conducted by the instructors and those conducted by the students in the course. The purpose of the instructor-led investigations was to model interpretation of data and to present exemplars of analyses of TAAS that the research team at SYRCE had conducted over the course of three years. The data investigations presented to the class came out of analyses done by the research team, many of which are detailed elsewhere (Confrey & Makar, in press), by teachers in the pilot study, or analyses that I conducted for the

purpose of the class. The student-led investigations were devised to provide the students in the course with shorter experiences in searching for data, open-ended exploratory data analysis, as well as structured opportunities to refine their analytic approaches to examining data.

Here is an example. In Unit 1 of the course, which focused on assessment, the preservice teachers worked in groups to design a lesson and assessment on subtraction of integers for sixth grade, to help them to start thinking about their teaching experience in Unit 2. They were then shown video excerpts of an innovative middle school algebra class which was covering a unit on subtraction of integers. In addition, the preservice teachers were given examples of student work and the item-level data from a pre- and posttest on subtraction taken by the sixth graders (they were also given a copy of the test). As an assignment, the teachers were asked to make observations about what the sixth grade students knew or had learned (or not learned) about subtraction by assessing student work (qualitatively) and exploring the data in Fathom (quantitatively). Observations were discussed the next class. In addition, the instructor demonstrated a number of innovative approaches to examining the data that provided additional insight into what the students had learned (see Appendix C).

Other class-led investigations looked at the relationships between student performance on the Iowa Test of Basic Skills (ITBS) and TAAS, comparing distributions highlighting gender differences of SAT scores and college grades, distributions of raw and scaled data from TAAS, linked distributions and distributions of the change in students' TAAS scores from 6^{th} to 7^{th} grade, and simulations of samples with varying sizes drawn from a population with a given passing rate. In addition, an optional workshop on conducting permutations tests (called scrambling in Fathom) as a means to compare groups was also given.

Student-led investigations included their own observations from the ITBS and TAAS data, analyzing their own pre-post test data from their teaching experience, exploring data available on the TEA website, a structured investigation of variation in passing rates through sampling from a population, and several investigations of data from exercises assigned in *Workshop Statistics: Discovery with Data and Fathom* (Rossman, Chance, & Lock, 2001).

3.4.3 Data Generation

This section briefly describes the data that was generated for this dissertation. Greater detail is provided in the results chapters. Quantitative data was collected in the form of a pre-post test of statistical concepts. Qualitative data consisted of class assignments, transcribed pre-post interviews, videotaped class discussions, final presentations and papers, and focus interviews.

Pre-post test

The pre-post test was designed to assess the teachers' understanding of statistics, with a particular focus on concepts of variation and distribution. The content of the pre-post test ranged from interpretation of graphs and appropriate descriptive measures, to conceptual questions about variation and distribution, to conceptual questions about the central limit theorem and sampling distributions. The questions were designed to include a wide range of difficulty levels to ensure that all students would have questions that they could answer, but also include difficult enough questions that growth would be able to be recorded. One formal question about confidence intervals was included to see if students who had already learned statistics would remember their formal content and to see how the students as a whole would perform on a traditional question when they were taught less formally.

The only difference between the pretest and posttest is that the pretest also collected demographic information as well as background in mathematics coursework, and asked the subjects to rate their facility with technology (email, internet, Excel, Fathom, and other statistical software). These questions were not asked again on the posttest, however the posttest also asked them to describe their previous statistical experiences and included a survey that assisted in evaluating the course. Because the posttest also served as the final exam, these "opinions" were expected to be rather positive and not necessarily reflect the prospective teachers' candid thoughts. These survey questions, therefore, were not analyzed.

To increase the validity of the statistics questions on the pre-post test, many of the questions were drawn from an assessment bank of items created by delMas, Garfield, and Chance (2001), the authors of *Tools for Teaching and Assessing Statistical Inference*, an NSF project at the University of Minnesota (www.gen.umn.edu/research/stat_tools). These items were designed for students in a

university statistics course that emphasized hypothesis testing, and some items that were more conceptual in nature were chosen because the course in the dissertation study was designed to touch on conceptual elements of hypothesis testing and sampling distributions. Additional items were taken from the Statistical Reasoning Inventory (Garfield, 2003), and a classic item (the Hospital Problem) assessing reasoning about sample size (Kahneman, Slovic, & Tversky, 1982). In addition, a pair of items were taken from a research study on distribution conducted by Pfannkuch & Brown (1996), which examined tolerance for variation in a probabilistic (dice) and real-world setting. The problem structure of these two questions was identical except for the context. It was of interest whether attention to context would change over the course of the study as Pfannkuch and Brown had found that students expect variation in probabilistic settings but interpret the same variability deterministically in a real-world setting. Many of the questions chosen for the pre-post test had been used in the pilot study successfully, increasing the confidence in the validity of these items.

As argued in the literature review, there is a lack of research on teachers' statistical reasoning in the context of assessment data, so there was no suitable instrument which probed statistical understanding in this context. Therefore, I developed several questions in the context of test data, using authentic data. I piloted these questions, along with the rest of the test to assess its length, with several graduate students in the department with varying backgrounds in statistics to be sure that questions were clear and assessed the intended concept.

A copy of the statistics portion of the posttest is included in Appendix D. The test consisted of 26 questions which fell into one of three major categories:

- Reading graphs-7 questions
- Variation and Distribution-11 questions
- Sampling Distributions-8 questions

18 questions were multiple choice, and 7 questions were free-response, and 1 was open-ended. Further details of particular items are given in the quantitative results chapter (Chapter 4).

Assignments

All written assignments were collected and photocopied (except for statistics exercise sets and their reflection on their model teaching) before being returned. A list of assignments is listed on the course syllabus (Appendix B). The major focus for the

dissertation was in examining how the prospective teachers' understanding and use of statistics interacted with their understanding and proclivity towards issues of equity. The only major assignment in which this interaction was systematically ensured was in the final inquiry project. The reflection papers, each focused on a particular "situation" regarding fairness in accountability were used as additional data to probe the teachers' articulation of their beliefs of equity further. Other assignments were kept on file, but used mostly for background information.

Inquiry Projects

The main assessment, and the capstone for the course, was the four-week inquiry project. The prospective teachers were given two weeks of class time to work on their investigation. In the third week, they presented the results of their investigation to the class (about 15 minutes plus questions) and received feedback from the instructors and their peers. The written portion of the project (12-15 pages in length) was due at the end of the fourth week. The assignment as it was given to the students in the course is in Appendix E. The layout of the written part of the project modeled a research paper or thesis: introduction, literature, method, results, and conclusions. In the introduction, the prospective teachers were to introduce their project, indicate how they chose it and why it was important, state a question they were trying to answer, and include a conjecture about what they would find. The second section, "Links to Equity", was meant to provide an opportunity for them to link the readings and discussions of equity that we had in the course to their investigation. This was also an opportunity for them to argue why their topic of inquiry was important from the standpoint of equity. Next, the method of analysis, including choice of data, was to be described. The fourth section, Results, was a place for them to provide details of their analysis and the results they found. Finally, the discussion and conclusion provided them with a chance to interpret their findings and link them back to their question and conjecture.

This final part of the course, focusing on the inquiry project, was designed to be the main source of data for the dissertation study. The reason for this is that (1) it was a longer project, providing the prospective teachers time to reflect on the link between their topic and equity, and the statistical evidence they would use to support their topic of inquiry; (2) it was more authentic than class assignments because the inquiry emerged from *their* interest; (3) it was a sufficiently complex task to be able to dig deeper into their understanding than a less complex task would permit. Finally, (4)

because of the design, students would present their results and then refine them, based on questions and comments from their peers, in a written paper. The researcher was then able to observe their thinking on two occasions: what they chose to unveil to their peers about their findings in a condensed timeframe, and how they elaborated on their thinking in a 15-page written paper. The condensed form informed the researcher on what they found most important or most interesting, as well as how they responded to questions and feedback, while the written form gave more insight into their approach to methodology and choice of evidence to support their findings.

Interviews

Individual interviews were conducted at the beginning and end of the course with all of the prospective teachers. Each interview was planned to last about 15 minutes in January and about 45 minutes in May. Interview questions were determined in advance (see Appendix G), but were not clinical, as they did not follow a prescribed script. As subjects responded, their thinking was probed to clarify and extend it. Using this approach, I intended not to compare precise responses but to gain insight into the teachers' thinking. The data from the interviews were analyzed qualitatively using a method based on Grounded Theory (Strauss & Corbin, 1998). The January interview consisted of a single task with several questions that sought to report the teachers' choice of language in describing and comparing distributions in the context of testing. The interviews in May also included this task as well as a semi-structured investigation in Fathom, described below and in Chapter 5.

For the task used in both the January and May interviews, a situation was described in which an urban middle school was trying to evaluate the effectiveness of a semester-long remediation course, called "Math Enrichment", intended for eighth-grade (14-year old) students believed to be "at risk" of failing the mathematics portion of the upcoming TAAS test. The decision for placement in the course was determined by the school guidance counselors; their criteria were unknown, but presumed to be based on the results of their 7^{th} grade math TAAS scores. Because the school was trying to decide if the remediation program was working, they were interested in how much the scores of each student improved and measured the difference in each student's score between his or her seventh grade math TLI and a practice test given near the end of eighth grade (with raw scores converted to TLI scaled scores). Each interviewee was shown a pair of dot plots (Figure 3.5) of authentic data with students who were in the math enrichment

class in the upper distribution and the rest of the students in the eighth grade in the lower distribution. It was emphasized that a positive value would indicate improvement in score and a negative would indicate a drop in score. Data points highlighted in red were those students classified as economically disadvantaged.

After describing the situation, teachers were asked to compare the relative improvement of students who were in the enrichment program with those who were not. The data and situation used in the interview were authentic and there was no obvious difference between the improvement of each group, as can be seen in the figure. Note also that the mean improvement in both those in the "enrichment" program and those who were not lies in the negative region. Rather than create hypothetical data that might allow for a clearer distinction in what the prospective teachers were focusing on, the researchers chose data from an actual school in a more complex and authentic setting—one that these teachers may be more likely to face in their own schools.

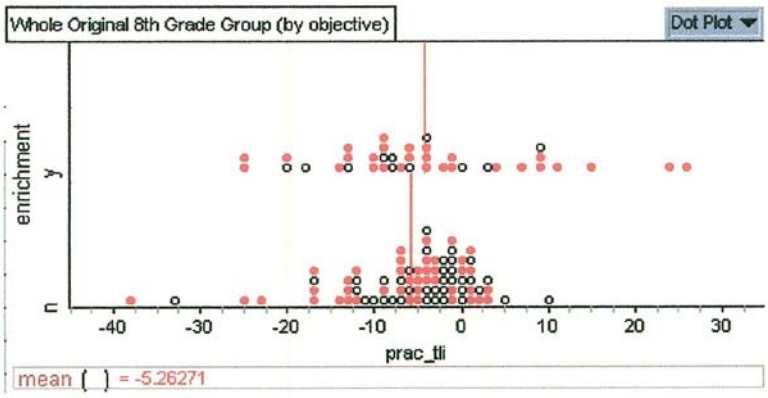


Figure 3.5: Graph shown to subjects during the interview task.

In all, twenty-two prospective teachers were interviewed in January (of which one interview was lost due to technical failure) and again in May (this time numbering eighteen due to attrition). Each interview was videotaped, transcribed, and coded to find the categories of concepts that would emerge from the data. The qualitative software NVivo (QSR, 1999) was used to assist in coding the data. General categories were sought initially through open coding to isolate concepts that might highlight thinking about variation and distribution, and those passages identified by these codes underwent finer coding resulting in eighteen preliminary categories. Since codes were not

predetermined, but rather allowed to emerge from the data, this portion of the analysis was not linear and underwent several iterations of coding, requiring a back-and-forth analysis as new codes were added, deleted or combined. Commonalities and differences were examined in passages coded under each node to better describe and isolate the category, determine dimensions and distinctions among participants' descriptions, and locate exemplars for each category.

In May, the same interview was conducted. In addition, the prospective teachers were asked to conduct a short investigation in Fathom. The Fathom investigations were designed to investigate how the prospective teachers would use dynamic statistical software in a more structured investigation. Although one cannot assume that the subjects' use of technology in a structured setting would parallel that in the open-ended projects where they had less technical support, it was hoped that the interviews might serve as a first approximation. Ideally, one would videotape and record all the actions and behaviors that the teachers used in conducting their three-week inquiry, but this was not practical. Therefore, the structured investigation was meant to serve as a way to describe ways that the teachers stated a conjecture and then used the software to search for evidence and develop a conclusion about a question that involved comparing the performance of two groups of students in the context of high-stakes testing. In addition to the regular video, their actions on the computer were recorded so that they could be later linked to the regular video, if needed. Details about the task are described in Chapter 5.

In addition to the January and May interviews, six students were chosen using stratified random sampling for additional "focus" interviews. The class was divided into a two-by-three matrix according to their performance on the pretest (low, middle, high) and their content focus (mathematics or science). One student was selected randomly from each of these six cells to capture a fairly representative slice of the class and to maximize the diversity of the responses. Each of the six prospective teachers were interviewed three additional times during the course, once during Unit 3, once as they were planning their inquiry projects, and once after they had completed the written final paper. These additional interviews were used to further probe thinking, particularly about the inquiry projects.

Class video

Each class meeting was recorded with two video cameras, one focused on the instructor and the other on the students in the course. These videos were created to keep track of, and later record, the content, activities, and discussion that occurred in the course, particularly since it deviated from the original syllabus during the semester. In addition, any discussions that occurred during the course, as well as the final presentations, were recorded. Only the tapes from the final presentation were transcribed for fine analysis.

3.4.4 Analysis of the Design Experiment

Because this study contains both quantitative and qualitative data, the analysis could be considered a mixed-method study. The pre-post test was analyzed primarily quantitatively, although one open-ended question (Q8) was analyzed qualitatively. A priori hypothesis tests were used to test conjectures that were predetermined (e.g. that there would be significant gains from pretest to posttest) or to test relationships between subgroups of students in the course that were hypothesized to show differences. These subgroups and the quantitative analyses conducted are explained in Chapter 4. As analysis was being conducted, some additional relationships were pondered and tested visually post hoc. Where interesting relationships were observed in the graphs, they were often subjected to a hypothesis test to further test the relationship, but these results are only speculative.

Most of the data from the study was analyzed qualitatively. The interviews and class presentations were transcribed and analyzed using the methodology of grounded theory (Strauss & Corbin, 1998). Under this methodology, the transcripts were first subjected to line-by-line open coding in the qualitative software *NVivo* (QSR, 1999) to create initial categories that would capture the phenomenon observed in the prospective teachers' own words and actions and to allow potential categories to emerge from the data that would need further investigation. Secondly, initial categories were organized into hierarchical trees to identify major categories and subcategories; during this process, the data were subjected to axial coding to begin to identify various dimensions of the categories and to search for common themes and processes among the larger categories. Next, the data were analyzed with selective coding to further investigate conditions and consequences, to better describe the phenomenon observed, and to create

a tentative theory that identifies and explains elements of the teaching practice. Because most of the analysis was done after the end of contact with the subjects, only a limited amount of theoretical sampling was conducted through additional interviews with six students that were chosen as a 'focus group' or by tapping into additional data sets (e.g. reflection papers).

3.5 SUMMARY

This chapter laid out the theoretical bases, design, and methodology of analysis of the study. The results are presented next, split into two chapters. Chapter 4 will report on the quantitative results of the pre-post test of statistical concepts, examining both the group performance as a whole and how particular subgroups of the prospective teachers performed on the test. This chapter will also report on the performance on the areas of the test that showed the greatest strengths, areas of difficulty, and areas of improvement in their performance with particular attention to key questions. Chapter 5 reports on qualitative results from interviews, reflection papers, and the prospective teachers' final inquiry projects. Chapter 5 also documents their articulation of concepts of variation and distribution in interviews, the behaviors they exhibited in using the software to conduct a semi-structured investigation, and the complexities that were uncovered in the process of conducting their final inquiry projects.

Chapter 4: Quantitative Results

This chapter will present the quantitative results of the study, which consists of results from the pre-post test. The purpose of the pre-post test in relation to the research questions was threefold: (1) To triangulate and inform the qualitative results—particularly understanding about variation and distribution, which will be presented in the next chapter; (2) to probe into specific areas and question types in which the prospective teachers excelled or struggled, (3) to examine patterns of learning among four particular subgroups of teachers based on: age, subject area, previous statistical experience, and ethnicity. The rationale for choosing these four subgroups will be discussed in Section 4.1.1. It should be emphasized that the purpose of the pre-post test was not to measure overall growth in the statistical content knowledge taught in the course, and therefore it was not aligned specifically to the course content. The reasons for this are threefold. For one, the purpose of this study is not to evaluate the content of the course. Secondly, the statistical content of the course was meant to be responsive to the discussions of equity in assessment that emerged throughout the course. Some areas of content, therefore, could not be planned in advance. As a result, some topics were taught in more depth and others less depth that originally planned as the course was modified throughout the semester. Finally, for the purpose of this dissertation, the focus was specifically on the prospective teachers' conceptions of variation and distribution. Therefore, for example, the test did not include the topic of linear regression even though the class spent over two weeks on this topic. The weight of importance of topics on the pre-post test was on assessing conceptual notions rather than on testing theoretically-based knowledge of procedures or theorems.

The pretest was administered on the first day of class and the posttest was given during the final exam period. Subjects were encouraged to do their best on the pretest, and the posttest counted for a small percentage of their overall grade (less than 10%). The test consisted of twenty-six content questions and was constructed to be a measure of the subjects' understanding in five areas: histograms (Q1-7); comparing groups (Q8); variation (Q11, Q21-23, 25); distribution (Q12, 24, 26ab, 27); and concepts related to the Central Limit Theorem (Q13-20). The pre- and posttest also included a section which asked subjects to rate their level of confidence, on a 5-point Likert scale, in each of the following statistical topics: descriptive statistics, statistical graphs, probability

distributions, sampling distributions and hypothesis testing. In addition, the pretest collected background on demographics and experience with technology and various software programs, including statistical packages and Fathom. The posttest additionally contained questions which were designed to inform the course evaluation, but were not analyzed for this study. A copy of the posttest is given in Appendix D.

4.1 OVERALL PERFORMANCE

The results of the content section of the test indicate that teachers made significant gains in the topics tested. 47.1% of the questions were answered correctly on the pretest and 68.9% on the posttest, with a mean gain of 5.7 questions ($s = 2.23$) out of 26 questions. This was a significant gain at the $\alpha = 0.05$ level ($t_{17} = 10.83$, $p < 0.0001$). Below is a scatter plot (Figure 4.1) of the number of questions each person correctly answered on the pretest (horizontal axis) and posttest (vertical axis). This representation, with line $y = x$ added, is helpful to begin to visualize patterns in improvement. Since all eighteen prospective teachers fall above the line $y = x$, we can see that all of them showed an improvement from pretest to posttest. The least-squares regression (LSR) line on the graph also indicates that there was a moderate correlation ($r = 0.54$) between their pretest and posttest scores; under the null hypothesis that there is no correlation, the likelihood of producing a correlation of 0.54 or higher is significant ($p = 0.022$). This indicates that the prospective teachers with higher scores on the pretest tended to have higher scores on the posttest.

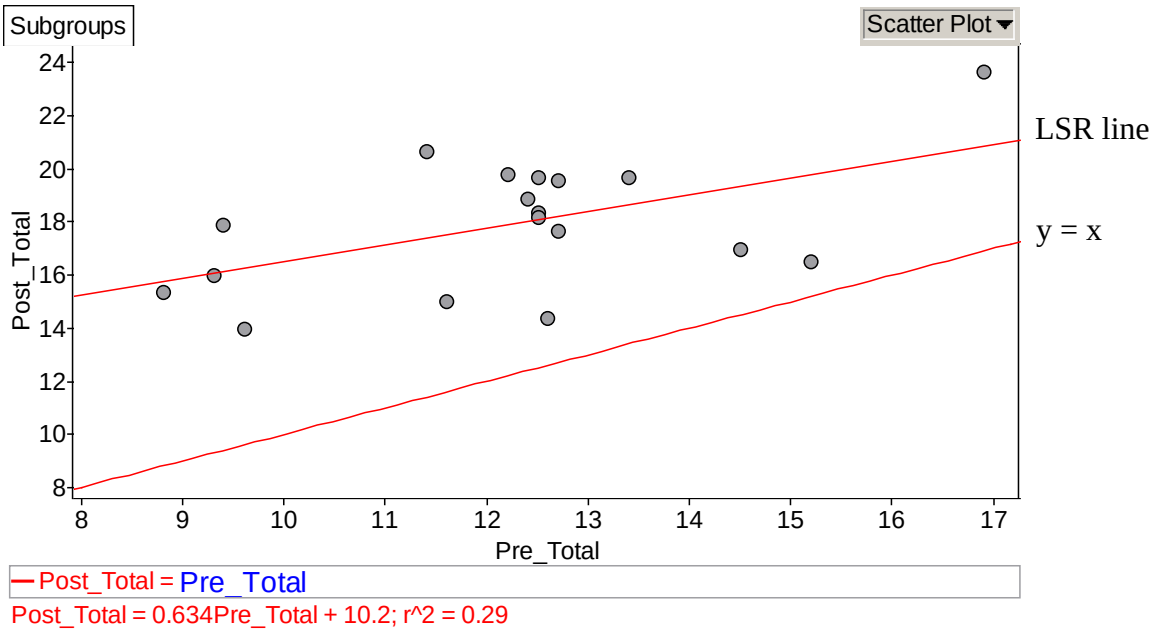


Figure 4.1: Overall performance of subjects on the pretest (horizontal axis) and posttest (vertical axis). The line y = x and the least squares regression line are shown to highlight improvement trends from pretest to posttest.

4.1.1 Subgroups

Although the number of subjects in the study was small (n = 18), it was of some interest to see whether there existed any differences in certain subgroups of prospective teachers in the class. For example, a fairly large proportion of the class (n = 7) consisted of minorities (African-American or Hispanic) and because of the emphasis in the course on equity, as a context to examine data, I wondered whether this context might have had a particularly positive effect on the minority students in the class. A good proportion of the class (n = 7) was also above the traditional college age of 18 – 22 years old, and there was some interest in whether more mature students might do better in such a complex environment. This comparison was particularly sought because of the research done by William G. Perry Jr. (1968/1999) and others (Evans, Forney, & Guido-DiBrito, 1998; King & Kitchener, 1994) that indicates that through the college years, students gain an increasingly less deterministic view of the world. The third subgroup that was investigated were those prospective teachers who had studied statistics in some form before or concurrently with this class (n = 11) in comparison to those who had not (n = 7), to see if the less-theoretical approach might act as an

equalizer on the posttest and to investigate whether prospective teachers who had already studied statistics would demonstrate an advantage coming into the course. Table 4.1 below summarizes the statistics backgrounds of the prospective teachers in the course. Note that the statistics experiences of the subjects in the study were very diverse, and only one had taken a formal, mathematical statistics course. Finally, it was conjectured that there might be a difference between those who were studying to be science teachers and those who were studying mathematics. It was conjectured that science teachers would likely have greater experience with data and feel more comfortable with analysis of graphs and basic statistics. On the other hand, theoretical statistics is better aligned with mathematics than science and so it was possible that prospective secondary mathematics teachers would have a stronger mathematics background and therefore might understand the underlying statistical concepts better. Gender differences were not compared because the number of males in the class was very small (n = 3) and it was not expected there would be any difference due to gender.

Table 4.1: Statistics background of prospective teachers in the study

Statistics Experience	Count
Mathematical Statistics Course (Statistics Department)	1
UT Probability (Mathematics Department)	2
UT Research Methods (taken concurrently with this study)	2
Applied Statistics Course (in a social science department)	3
Statistics within mathematics or science coursework	2
Other (Statistics courses taken over 20 years ago)	1
None	7

To check for possible confounding between subgroups, I examined the six pairwise interaction patterns for the four categories of subgroups (Table 4.2). Sample size was too small to use a Chi-Square test for independence between subgroup categories. A Fisher Exact Test indicated that there was not enough evidence to reject a null hypothesis of independence between pairs of subgroups, but again, the size of the sample (n = 18) would make significance very difficult to obtain. In observing intersections between subgroups, there is a possibility of confounding between ethnicity, age, and statistical experience:

- 14% of the minority students (1 of 7) were over 22 years old, compared to 55% of the Caucasian students (6 of 11).
- 43% of the minority students (3 of 7) had taken statistics by the end of the course, compared to 73% of the Caucasian students (8 of 11). 29% of the minority students (2 of 7) had taken statistics previous to this course, compared to 64% of the Caucasian students (7 of 11).
- 71% of the students over 22 (5 of 7) had taken statistics by the end of the course, compared to 55% of those who were younger (6 of 11).

Table 4.2: Pairwise intersections of subgroups to check for possible confounding between subgroups.

		Minority		Over 22		Subject	
		N	Y	N	Y	Math	Science
Over 22	N	5 (28%)	6 (33%)	-	-	-	-
	Y	6 (33%)	1 (6%)	-	-	-	-
Subject	Math	6 (33%)	3 (17%)	5 (28%)	4 (22%)	-	-
	Science	5 (28%)	4 (22%)	6 (33%)	3 (17%)	-	-
AnyStats	N	3 (17%)	4 (22%)	5 (28%)	2 (11%)	4 (22%)	3 (17%)
	Y	8 (44%)	3 (17%)	6 (33%)	5 (28%)	5 (28%)	6 (33%)

Student responses were compared in each of the four categories above with two-sample t-tests (with unpooled variance); significant results were also visually examined with a graphical representation. Analyses were carried out in Fathom™. Because of a concern that multiple comparisons of the data set may lead to too high of a family-wise Type I error rate, the alpha level was reduced to a more conservative level from $\alpha = 0.05$ to $\alpha = 0.01$ for each comparison using Dunn's method (Lomax, 2001). Below (Table 4.3) are the results of the pretest and posttest (and improvement from pretest to posttest) for each of the four subgroups, as well as the class overall. Only students who took both the pretest and posttest were counted (n = 18). Cells that are highlighted indicate that there was a significant difference between these groups.

Table 4.3: Mean (standard deviation) number of problems correct on the pretest, posttest, and change (from pretest to posttest are given), with comparisons (and p-values) for four pairs of subgroups. Subgroups with means that are significantly different ($p < 0.01$, unpooled variance) are highlighted.

Pre- and Posttest Results		Mean Number Correct (standard deviation) and p-values (out of 26 Questions)					
Subgroup		Pretest	p	Posttest	p	Change	p
Minority	Yes (n=7)	11.3 (2.08)	0.13	**16.1 (1.87)**	**<.01**	4.81 (2.46)	.22
	No (n=11)	12.8 (1.94)		**19.1 (2.12)**		6.25 (1.99)	
Under 23	Yes (n=11)	12.1 (2.03)	0.72	17.1 (1.83)	.12	5.01 (2.29)	.09
	No (n=7)	12.5 (2.33)		19.2 (2.92)		6.76 (1.77)	
Any Stats	Yes (n=11)	12.3 (2.10)	0.94	**19.2 (2.07)**	**<.01**	**6.95 (1.21)**	**<.01**
	No (n=7)	12.2 (2.24)		**15.9 (1.55)**		**3.71 (2.05)**	
Content Area	Math (n=9)	12.7 (2.54)	0.37	18.2 (2.81)	.66	5.50 (2.99)	.73
	Science (n=9)	11.8 (1.54)		17.7 (2.24)		5.88 (1.24)	
Overall	n = 18	12.2 (2.09)	n/a	17.9 (2.48)	n/a	**5.69 (2.23)**	**<.01**

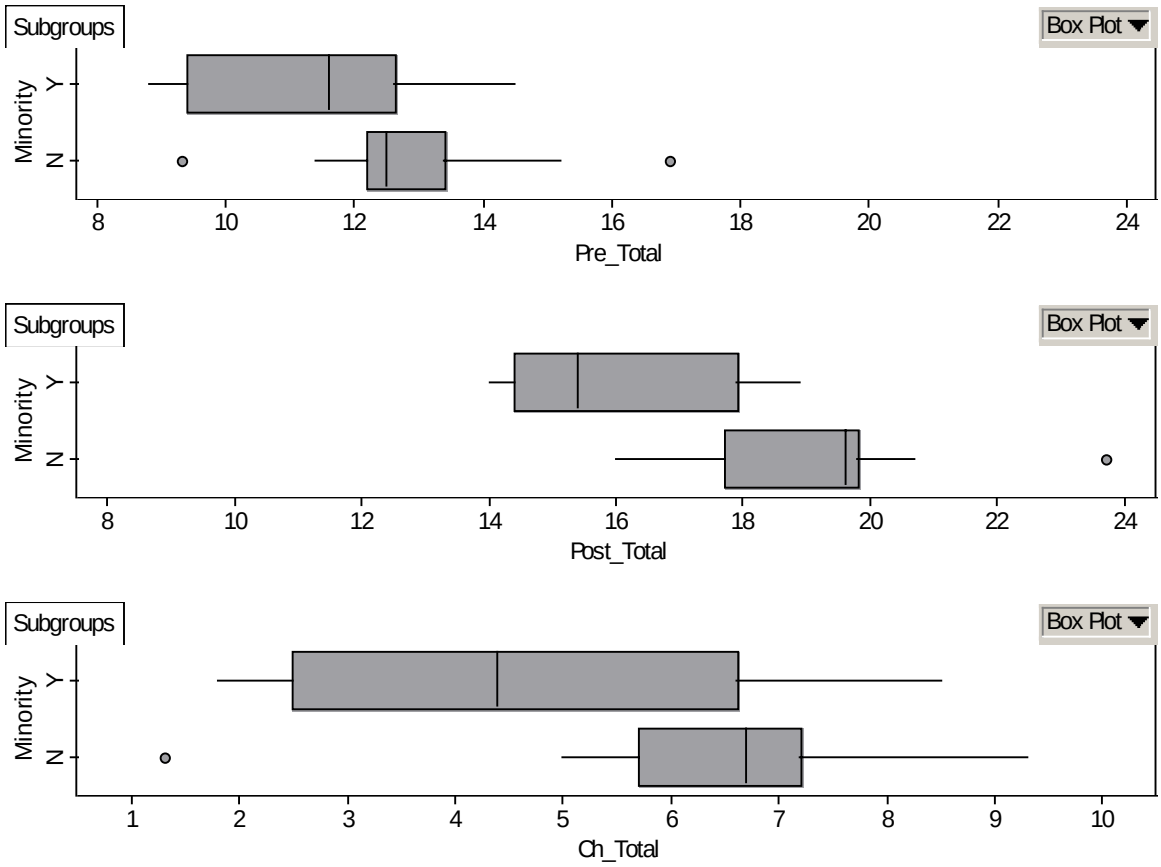


Figure 4.2: Boxplots of the pretest scores (above), posttest scores (middle) and improvement in scores (bottom) between minority (African-American and Hispanic, n = 7) and non-minority (Caucasian, n = 11) subjects, from Table 4.3.

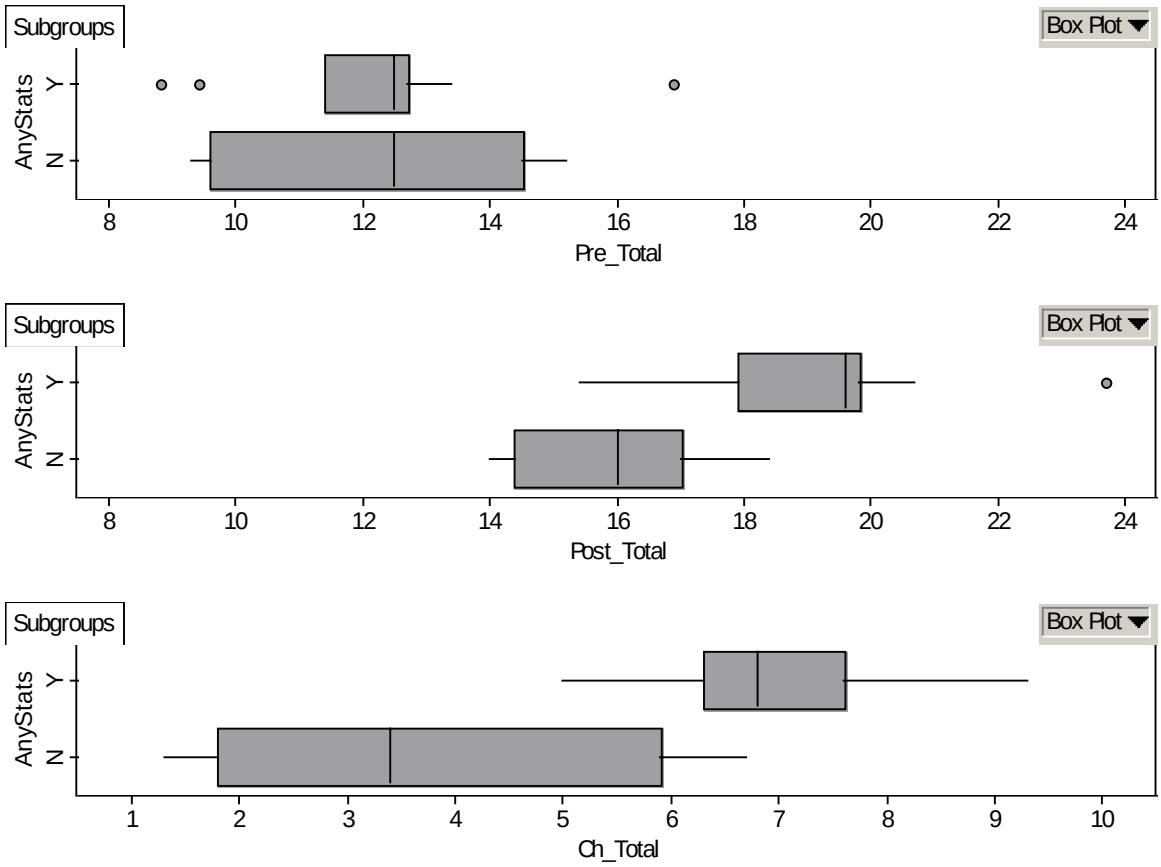


Figure 4.3: Boxplots of the pretest scores (above), posttest scores (middle) and improvement in scores (bottom) between those students with (n = 11) and without (n = 7) outside statistical experience from Table 4.3.

From Table 4.3 and Figures 4.2 and 4.3 above, a few observations can be made. For example, one can see that although there was little difference in the number of questions answered correctly on the pretest between those prospective teachers who had statistical experience and those who didn't, there was a significant difference in the improvement between these two groups. These results are the opposite of what was anticipated: the researcher had conjectured that the teachers who had previously studied statistics would probably show a significant advantage on the pretest over those who had not, but thought that the course might tend to equalize this difference. Instead, it appears that previous experience in statistics did not appear to benefit teachers on the questions coming into the course, but perhaps the extensive experience in data analysis during the course tended to solidify and clarify theoretical work which was not well

understood in previous courses, resulting in greater improvement. These results may indicate that although theoretical work does not, on its own, tend to increase understanding of concepts of variation and distribution, theoretical work combined with experience in exploratory data analysis work may provide a greater understanding of variation and distribution than either of these two experiences alone. Another possible explanation would be that those who had previously studied statistics were able to "review" previous statistical concepts during the course that they had forgotten, explaining the higher improvement by those prospective teachers who had previously studied statistics.

Another result is that although the minority teachers in the course did not have significantly different mean scores on the pretest, their mean posttest scores did differ from the Caucasian teachers significantly. This is a concern, although this may be due to confounding between statistical experience and ethnicity. Possible confounding with statistical experience is particularly an issue because of the significantly higher posttest scores for those with statistical experience; only 43% of the minority students (African-American and Hispanic) in the course had statistical experience compared to 73% of the Caucasians. Note also that the overall improvement in statistical content knowledge for the class was significant ($p < 0.01$).

4.1.2 Confidence

Confidence is a critical issue in addressing the participation of under-represented groups in mathematics and science, particularly women and minority students. The literature on gender points out that a significant predictor of participation by women in upper level mathematics and science courses is whether "they express less confidence than males in their abilities to master mathematics and science in careers requiring these skills than in other subject areas and careers" (p. 250, Lynch, 2000 citing Eccles, 1995). Similar results for minority students have been found (Seymour & Hewitt, 1997). Because of the importance of confidence, it was also of interest whether there were possible differences for the four pairs of subgroups in their reporting of confidence in their understanding of statistics. The preservice teachers rated their own comfort level in statistics in five different topics (descriptive statistics, statistical graphs, probability distributions, sampling distributions, and hypothesis testing) at the beginning and end of the course, using a 5-point Likert scale. The results are given below in Table 4.4 and Figure 4.4. Again, cells that are highlighted in the table indicate significant differences

in comfort levels for each of the four subgroups explored earlier in Subsection 4.1.1: ethnicity, age, statistics experience, and content area. For the subgroup comparisons, as before, the significance level is lowered to $\alpha = 0.01$ to guard against the risk of family-wise Type I error rate.

Table 4.4: Mean (standard deviation) confidence level in five topics of statistical content as self-reported by subjects before and after the course, split by four pairs of subgroups and based on a five-point Likert scale. In addition, p-values and the overall mean change are reported.

Pre- and Posttest Confidence Results		Mean (standard deviation) self-reported confidence levels and p-values					
Subgroup		Pretest	p	Posttest	p	Change	p
Minority	Yes (n=7)	**1.80 (0.53)**	**<.01**	3.06 (0.65)	0.13	1.26 (0.66)	0.41
	No (n=11)	**2.56 (0.46)**		3.53 (0.48)		0.97 (0.75)	
Under 23	Yes (n=11)	2.25 (0.57)	0.94	3.32 (0.40)	0.86	1.07 (0.63)	0.92
	No (n=7)	2.28 (0.70)		3.39 (0.84)		1.11 (0.87)	
Any Stats	Yes (n=11)	2.46 (0.58)	0.07	3.45 (0.39)	0.44	0.99 (0.72)	0.50
	No (n=7)	1.95 (0.54)		3.18 (0.82)		1.23 (0.73)	
Content Area	Math (n=9)	2.16 (0.64)	0.49	3.51 (0.52)	0.26	1.35 (0.80)	0.12
	Science (n=9)	2.37 (0.59)		3.19 (0.63)		0.83 (0.53)	
Overall	n = 18	2.26 (0.61)	n/a	3.35 (0.58)	n/a	**1.09 (0.71)**	**<.01**

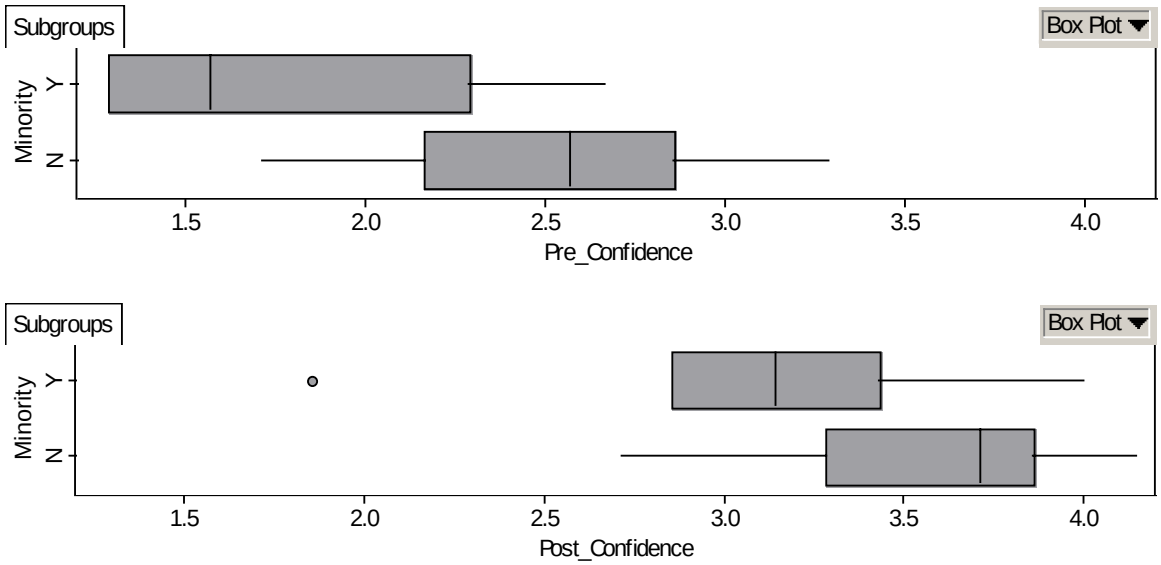


Figure 4.4: Box plot of significant mean difference in levels of personal confidence in understanding of statistical content on the pretest between minority and Caucasian students as shown in Table 4.4.

The results show that for subgroup pairs comparing age, statistical experience, or content area, the mean difference in confidence levels was not significant. Minority teachers in the course, however, did show significantly lower mean confidence in their knowledge of statistics than Caucasian teachers when coming *into* the course, despite the fact that their knowledge of statistics was not significantly lower (Table 4.3). Because there was no significant difference in mean confidence levels (before or after the course) for those who had previously studied statistics, it is unlikely that this result (like that for posttest performance) is due to confounding between minority students and those who had previously studied statistics. This result is particularly troubling given that confidence can influence their perceptions about their potential for success (Seymour & Hewitt, 1997), and hence decisions about continuing coursework in mathematics and science. The fact that these students are already majors in mathematics or science, and so likely have higher levels of confidence then their peers who are not math or science majors, makes this result all the more discouraging.

The results at the end of the course, however, are more encouraging. That the gap in confidence between minorities and Caucasians was softened during the course is heartening and may point to the need for more opportunities for minority students to engage in experiences similar to what they had in this course. This result may indicate

that the heavy emphasis on equity in the course gave the minority students a greater sense of confidence in doing statistics within that context, since they could more easily relate to problems faced by minorities in schools. Another conjecture is that the reform-based approach to learning in the course was more compatible to the learning styles of the minority students than their previous experiences with these topics in their mathematics, science, or statistics coursework. In Chapter 5 (Qualitative Results), I will discuss the inquiry projects, in which many of the minority students were actively engaged and empowered by the opportunity to investigate a problem compelling to them.

Also note that the overall improvement in subjects' confidence in statistical content knowledge for the class was significant ($p < 0.01$), with a gain of about 1.1 points ($s = 0.71$) on a 5-point scale. Although not significant, it is interesting that the mathematics preservice teachers gained more confidence in the course than their science teacher counterparts. In fact, the science preservice teachers entered the course with a higher level of confidence in statistics than their mathematics counterparts, but at the end of the course this trend was reversed. It is possible that the mathematics students found the statistical content more interesting, or that they recognized that they would likely teach statistics in their mathematics courses. During the 3-day teaching experience, for example, the science students complained that they were being asked to teach scatterplots and association of variables to their biology classes as they did not see the relevance of this topic in biology. The mathematics preservice teachers, however, appeared to enjoy teaching the topic and recognized that it was already part of the Algebra 1 curriculum.

4.1.3 Interaction between performance and confidence

At least a moderately positive correlation might be expected between comfort level with the statistical content tested and performance on the test, but this was found not to be the case on either the pretest ($r = .33$, $p = .19$) or posttest ($r = .26$, $p = .29$). Also note that the overall improvement in subjects' confidence in statistical content knowledge for the class was significant ($p < 0.01$), with a gain of about 1.1 points ($s = 0.71$) on a 5-point scale. Although not significant, it is interesting that the mathematics preservice teachers gained more confidence in the course than their science teacher counterparts. In fact, the science preservice teachers *entered* the course with a higher level of confidence in statistics than their mathematics counterparts, but at the end of

the course this trend was reversed. The change in order of confidence between the prospective mathematics and prospective science teachers could be due to chance, particularly since the degree of difference in confidence between these two groups was not significant. It is also possible that the mathematics students found the statistical content more interesting, or that they recognized that they would likely teach statistics in their mathematics courses. During the 3-day teaching experience, for example, the science students complained that they were being asked to teach scatterplots and association of variables to their biology classes as they did not see the relevance of this topic in biology. The mathematics preservice teachers, however, appeared to enjoy teaching the topic and recognized that it was already part of the Algebra 1 curriculum at the school where they conducted their practice teaching.

4.2 PATTERNS OF PERFORMANCE ON TEST QUESTIONS

Overall, performance on the 26 questions evaluated for this study showed a great deal of variation in performance, both on the pretest and posttest. Both tests contained questions within the entire range of performance, from 0% to 100% correct. Eighteen questions on the tests could be considered multiple choice, with eight providing two to three answer choices and ten providing four to five answer choices. These eighteen questions were all scored dichotomously as correct (1) or incorrect (0). Four of the remaining eight questions were short response questions and four were more open-ended questions. Both of these types were scored on a sliding scale with partial credit possible. The scatter plot below (Figure 4.5) represents the proportion correct for each question on the pre-post test, with the residual plot highlighting the improvement of each question from pretest to posttest. The line $y = x$ was drawn to emphasize questions that displayed particular improvement and to form the basis for the residual plot. Five question types were also identified according to topic, as described in the introduction to this chapter, with the legend indicating the topic tested by each question.

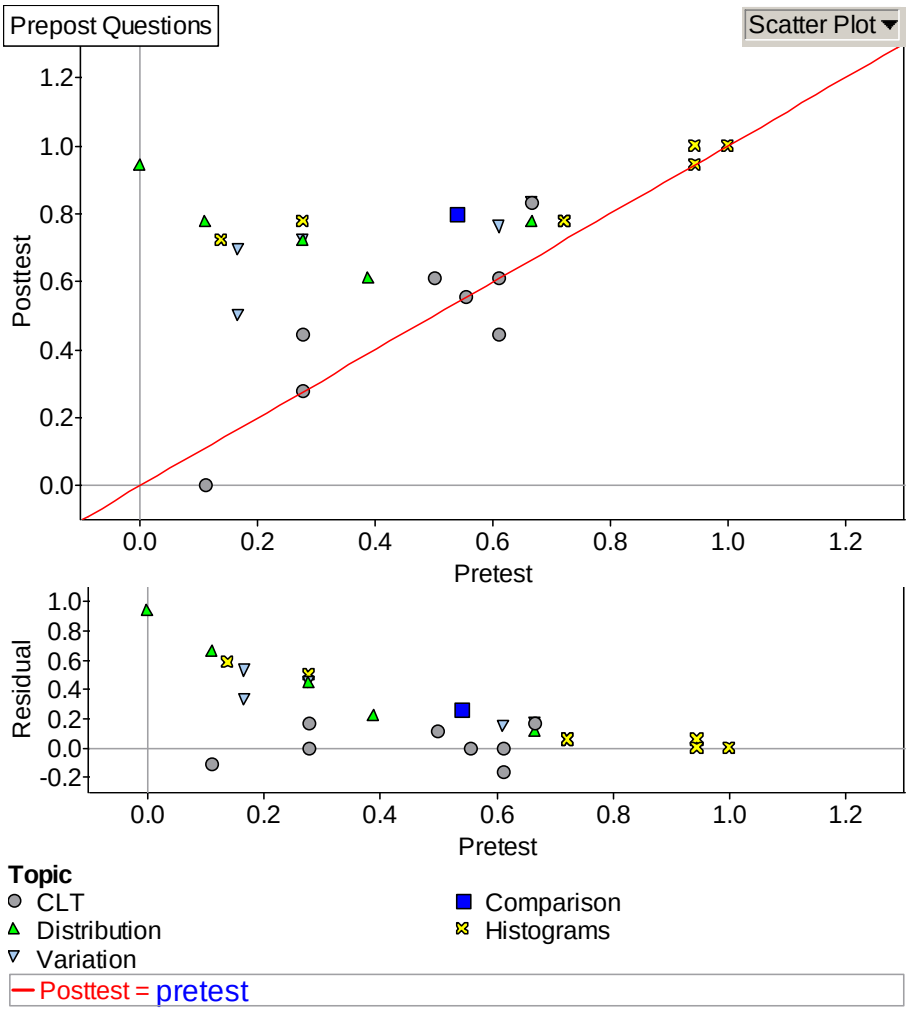


Figure 4.5: Patterns of performance on the 26 questions on the pretest (horizontal axis) and posttest (vertical axis) with the legend indicating the topic tested and the line y = x and residual plot shown to highlight questions that showed improvement.

From the scatter plot above (Figure 4.5), one might make the following observations. For one, the graph shows that the easiest questions on the pretest also showed high levels of performance on the posttest, as might be expected. For this reason, these questions also show little or no improvement in performance when one examines the residual plot. As stated above, the pre-post test included conceptual questions from five different categories: Histograms, comparing groups, variation, distribution, and the Central Limit Theorem. Table 4.5 below displays the performance on the pretest and posttest (and improvement) in each of these question categories. One

other type of question also emerges as showing no overall improvement from pretest to posttest, those questions which tested understanding of the Central Limit Theorem (Q13 – Q20). While students had some experience with simulations involving sampling distributions near the end of the course, time constraints due to schedule adjustments earlier in the course did not allow for formalization of the Central Limit Theorem as was initially intended. The table also shows that the two concepts that were the weakest on the pretest (variation and distribution) also posted the greatest gains on the posttest. This may be due to regression towards the mean, although since these two concepts were given greater emphasis during the course, it is likely a result of the treatment.

Table 4.5: Performance on five categories of questions on the pre-post test

	N	Pretest	Posttest	Change	
		Mean (sd), percent correct	Mean (sd), percent correct	Mean (sd), percent correct	p-value
Histograms	7	4.8 (1.33), 68%	6.0 (1.08), 86%	1.3 (1.70), 18%	<0.01
Comparing Groups	1	0.5 (0.18), 54%	0.8 (0.12), 80%	0.3 (0.19), 26%	< 0.01
Variation	5	1.9 (0.96), 38%	3.5 (1.15), 70%	1.6 (1.35), 32%	<0.01
Distribution	5	1.4 (0.98), 29%	3.8 (0.99), 77%	2.4 (1.20), 48%	<0.01
Central Limit Theorem (CLT)	8	3.6 (1.38), 45%	3.8 (1.56), 47%	0.2 (1.98), 2%	0.74
Overall	26	12.2 (2.09), 47%	17.9 (2.48), 69%	5.7 (2.23), 22%	<0.01
Overall without CLT	18	8.6 (1.95), 48%	14.1 (2.10), 79%	5.5 (1.71), 31%	<0.01

4.2.1 Areas of strength

It is worth noting particular areas of strength shown by the prospective teachers on the pretest and posttest. Areas of strength were identified in a normative fashion with performance at or above the third quartile of performance of the 26 questions on the pre-post test. The first five questions on the test (Q1-5, adapted from delMas et al.,

2001) were ones which showed the highest level of performance on the pretest, 72% - 100% correct. These questions dealt with simple interpretation of histograms, testing whether the teachers could identify what was represented on each axis and correctly make use of the frequency information provided by the histogram. These same five problems were also identified with high performance on the posttest with 78% - 100% of students answering them correctly.

Three other questions also showed high performance on the pretest, all scoring 67% correct:

1. the Dice problem (Question 11, adapted from Pfannkuch & Brown, 1996);
2. the identification of a point in a distribution of a sample versus one in a sampling distribution of means (Question 12, adapted from delMas et al., 2001); and
3. a dichotomous question on the shape of a sampling distribution of means for n = 25 (Question 14, adapted from delMas et al., 2001).

Performance on the dice problem is quite interesting and will be discussed in Subsection 4.2.7. The other two, although both dealt with sampling distributions (identified as the weakest area of performance, see Table 4.5), are not that surprising once one looks beyond the general topic. Question 12, for example, asks students to interpret the meaning of a single point in a distribution of a sample and its sampling distribution. In practice, this has been identified as an area of extreme difficulty (see for example, Makar & Confrey, in press), but this particular question states the information needed to answer the question if read carefully. The final question was surprising, but given that a "guess" on this question would yield an expected performance of 50% correct, the performance on this question is not so surprising.

On the posttest, seven questions were found at or above the third quartile, measuring at least 80% correct. All of the questions discussed above as areas of strength on the pretest showed also strong performance on the posttest, scoring at least 78% correct. Two additional questions appeared at or above the third quartile of performance of all questions evaluated on the test: Question 8 (discussed in Subsection 4.2.4) and Question 26a, which dealt with the identifying a skewed distribution as skewed left or skewed right. Statistics students frequently say that the determination of a skewed left distribution as one with a longer tail on the left side as counterintuitive. While a very high percentage of students answered this question correctly on the posttest (94%) not a

single student had answered it correctly on the pretest. Given that half of the students in the course had studied statistics before the course began (two students took statistics concurrently with this course), this might be surprising. However, consider that almost all of the students that had studied statistics previously did so in a course in which they likely focused on normal distributions. Such high improvement on this question was likely because scores of basic skills tests are frequently skewed left (because of a ceiling effect), so students in the course would have had a great deal of experience with skewed distributions in their work with testing data.

4.2.2 Difficulties

In addition to strengths, the pre-post tests also highlighted particular areas where the prospective teachers struggled. An overview of difficult questions will be given here with these areas further discussed in the next subsection.

In general, consistent areas in which the preservice teachers found difficulty on the pretest were basic definitions and properties of measures of center (e.g. in skewed distributions or histograms) and intuitions about distribution of small samples in *strongly contextual* problems. Difficult areas on the pretest were defined by performance in approximately the bottom quartile of all questions, or less than 30% correct. Four questions on the test probed the prospective teachers' understanding of properties of means and medians in skewed distributions (Questions 24, 26a, 26b, and 27). On the pretest, performance on these questions ranged from 0 – 39% correct. This was surprising given that about half of the students had already had introductory statistical training but again may indicate that their previous study of statistics emphasized normal distributions. Another very poor area of performance on the pretest involved estimating the median and mean of univariate data in a histogram (Questions 6, 7) with 28% and 14% of students answering correctly, respectively. Many of the teachers failed to include the frequency of the bins in their estimation and simply found the mean or median of the centers of each bin on the horizontal axis. This may point to their inexperience in handling data distributions. Most were able to correctly identify the meaning of the horizontal and vertical axes of a histogram as well as use the frequencies of the bins to count the size of the sample, so it was clear that they had seen and worked with the representation before, but their neglect of using the frequency to calculate the median of the data, for example, may indicate that their understanding of histograms was superficial. A third area in which difficulty was noticed was in

interpreting strongly contextual problems probing intuition about variation in small samples (Questions 22, 23, and 25). The particular difficulty with these questions is discussed in greater detail below. Three additional questions, all assessing conceptual understanding of the Central Limit Theorem, were also among the areas of difficulty on the pretest. As stated above, sampling distributions are particularly difficult concepts; all three of these questions were multiple-choice questions, but unlike Question 14 with two answer choices, these each had five answer choices.

At the end of the course, the areas in which students indicated difficulties had changed for the most part. Because there was overall growth in understanding, "difficulty" on the posttest was measured with a somewhat higher standard but still at or below the first quartile located at 61%. All but two questions at or below 61% correct involved understanding of the Central Limit Theorem. Two other questions on the posttest, both of which 61% of students answered correctly, were Question 25 (discussed in Section 4.2.7) involving identifying a need to use variation in an open-ended question concerning the distribution of categorical data in a small sample, and Question 27, which asked students to estimate the proportion of scores above the mean in a skewed distribution. Performance on Question 25 (50% of students answered it correctly on the posttest), was lower than the other three open-ended questions on the test, which produced outcomes of 69% - 83% correct. In the case of Question 27, which was multiple-choice, students were asked how many scores would be above the mean: more than half, less than half, about half, or can't tell. Interestingly, they performed slightly better on a different question in which they were asked whether the median would be above the mean in a left skewed distribution, answering 72% correct (Question 26b); the related concept stated less formally seemed to induce a slightly lower performance compared with the more formal property of means and medians in a skewed distribution. The performance on these two questions, however, is close enough that one cannot say the difference is meaningful.

4.2.3 Areas of growth

Questions that showed the most growth (at least 40% improvement) were Questions 6, 7, 22, 23, 24, and 26a,b. All of these questions indicated improvement at or above the third quartile, with a 44%-94% increase in percent of students answering correctly from pretest to posttest. The majority of these questions were in the topic of basic statistical concepts. For example, Questions 6 and 7 asked students to use a

histogram to estimate the mean and median of a distribution. On the pretest, many students had simply ignored the frequencies of the bins even though they correctly identified the y-axis as showing frequency in an earlier question. The growth in this area indicates that these measures of center may hold more meaning for them than at the beginning of the course.

Even though the other questions with high improvement (Questions 22, 23, 24, and 26a,b) tended to fall at the end of the test, there did not seem to be any issues related to fatigue as the last nine questions (last third of the test) generated reasonably similar results ($\bar{x}$ = 71%) to the overall mean of 69% on the posttest. Questions 24, 26a, and 26b were also basic statistical concept questions that probed into students' skill in vocabulary and characteristics of non-symmetric distributions. Question 24, for example, asked students to identify which measure of center and spread would be most appropriate for a distribution displaying outliers. Questions 26a and 26b were concerned with vocabulary regarding distribution shape and relative positions of the mean and median in a skewed distribution. Once again, all three of these questions may be an indication that these students had insufficient previous experience with distributions that are not normal.

Questions 22 and 23 were key questions that sought to probe the prospective teachers' intuition about variation. Here, I will briefly discuss Question 22. Questions 23 (Hospital problem) is discussed in Section 4.2.6. Question 22 is stated below:

22. 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.

In this Question, it is necessary to consider not only the difference between 90 degrees and the typical temperature in each city, but something about the variation in temperatures to get a sense of how unusual a 90 degree temperature would be within the distribution of temperatures for each city. On the pretest, only 17% of students recognized a need to consider more than just average temperature in each city, while this figure jumped to 69% on the posttest. This was encouraging as it could be interpreted that a higher percentage of the preservice teachers recognized the need to consider variation when making judgments about data. This, together with the growth shown in building intuition about variation of small samples shown in Question 23,

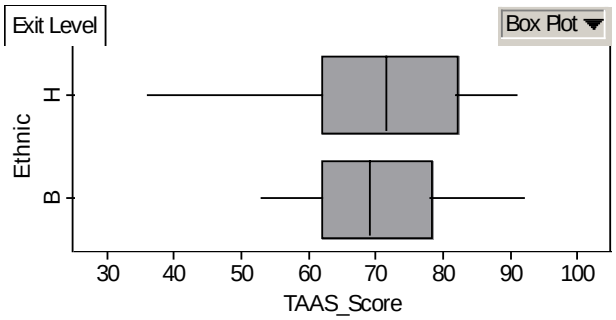
points to an overall improvement in understanding of subtleties of variation in the context of an everyday application that non-statisticians might encounter.

In the next few sub-sections, I will discuss performance on a few particular problems on the test which demonstrated interesting patterns of performance.

4.2.4 Question 8 (Comparing Groups)

I developed Question 8, shown below, using a sample of authentic student-level data of TAAS performance from a local urban school district. The idea for this question came from a study by Pfannkuch and Brown (1996) in which university students interpreted a very small variation in the difference in median performance on two tests, as represented on box plots, as meaningful even though the students interpreted similar differences for data of coin flips to be due to variation. Experiences at SYRCE (Confrey, in preparation) had also indicated that schools, in interpreting their students' test data, often put too much emphasis on summary statistics without regard for either the distribution of data or the size of the sample. Therefore this question was developed to assess what kinds of information teachers would pay attention to when comparing the performance of two groups of students using data presented in both graphical and summary form. Teachers' choice of language to articulate variation and distribution in comparison of groups is further assessed in the qualitative results presented in Chapter 5.

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.



Group	Number of Students	Mean TAAS score	Percent Passing TAAS
African-American	31	71.0	48.4
Hispanic	228	71.5	61.4

8. List at least 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…"

Each response given by students was first typed into a spreadsheet with the student's name and whether it originated from the pretest or posttest. Next, all of the responses were combined without this information so that they could be categorized by the type of response given. From the list of 118 responses, 17 initial categories were created into which the responses were sorted. These categories were then ordered by level of statistical complexity and rated on a scale of 0 – 5 according to the following rubric (Table 4.6):

Table 4.6: Responses and rubric describing each level of statistical responses given in Question 8.

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[2]." -"There are much lower scores in the lower quartile for Hispanic students."

[2] Although this response may indicate that students are still developing their understanding of the relationship between variability and sample size, I was interested here in whether students indicated a need to consider variability in making comparisons rather than whether the notion they had was correct.

After each response was coded, it was matched back with its subject. From the three responses of each student on the pretest and posttest, only the scores of the two highest-scoring responses were added for a total score for each subject's test (with a maximum of 10 points) to emphasize their best responses and to minimize the effect of fewer than three responses sometimes given on the pretest. These scores are used as a measure to determine if a higher level of statistical complexity is observed by the subjects on the posttest compared to the pretest. While there was a significant improvement ($\bar{x}$ = 2.7, s = 1.85, $p < 0.01$ or mean expected gain of 1.6 to 3.7 levels at the 95% confidence level) from pretest ($\bar{x}$ = 5.3, s = 1.81) to posttest ($\bar{x}$ = 8.0, s = 1.24), there was only a weak correlation between performance on the pretest and that of the posttest (r = 0.31). This indicates that students coming into the course with less experience in describing distributions and comparison were at no disadvantage on the posttest. On the other hand, the correlation between the pretest and improvement was strongly negative (r = -0.77) indicating that there may have been a ceiling effect (Figure 4.6). A modified version of the course described in this dissertation was also conducted the following semester at Washington University in St. Louis with a mixture of graduate and undergraduate students and this same question was used on the pre- and posttest in that course with similar results (Confrey et al., 2004).

The results of this question indicate that teachers entering the course paid attention to more simplistic characteristics of the two groups when making comparisons. For example, 62% of responses on the pretest rated at a level 2 or lower, indicating that most responses were superficial at best, stating no more than basic information given in the table. This improved on the posttest, where only 23% of responses were recorded at level 2 or lower. On the higher end, only 21% of responses on the pretest were at level 4 or higher, indicating that a low percentage of responses made use of statistical information at the beginning of the course. This percentage jumped to 47% on the posttest when nearly half of all responses used statistical information. Recall that each student was asked to provide three conclusions/responses, so while 47% of *responses* were at a level 4 or 5, 100% of the *students* gave responses at levels 4 and 5 on the posttest compared to 44% of students on the pretest.

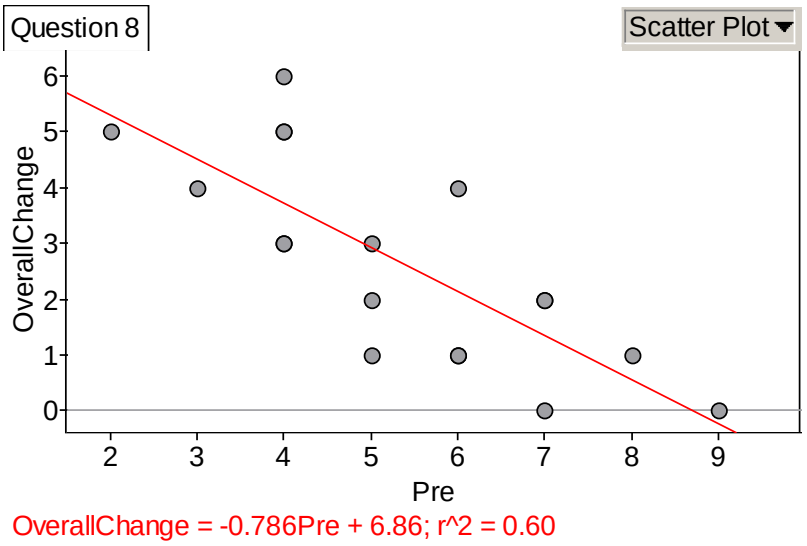
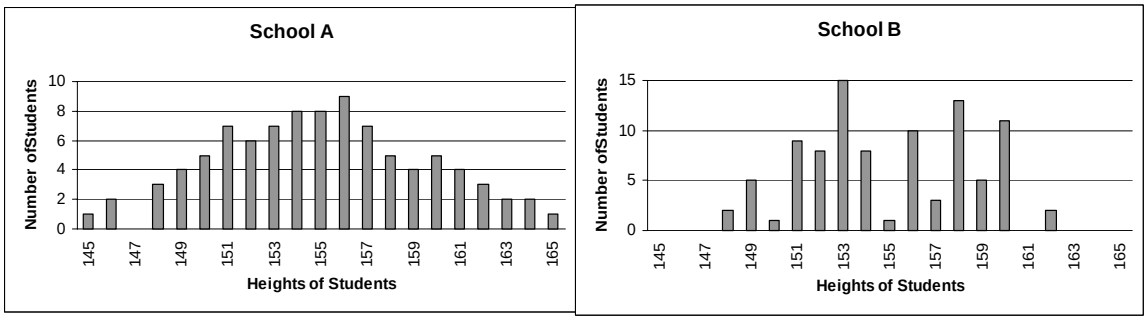


Figure 4.6: Association between pretest and improvement on levels of responses in Question 8.

4.2.5 Question 21 (Variability as bumpiness)

One common problem that novice students have is in understanding the word "variability" with respect to a distribution. Rather than variability in a distribution as a characteristic of "spread", Rossman, Chance, and Lock (p. 109, 2001) noted that frequently the notion of variability is interpreted as a distribution either having a greater number of distinct values or as appearing "bumpier" with respect to its shape. In Question 21, students were asked the following question, adapted from delMas et al. (2001):



21. The graphs above describe some data collected about Grade 7 students' heights in two different schools. Which graph shows more variability in students' heights? Explain why you think this.

The first histogram shows a somewhat normal distribution with a larger spread of scores, representing data with more statistical variation, but also capturing a possible conceptual confusion that variability was only measured by the range of scores in the data or that variability meant a distribution had a greater number of distinct values in the data. The second graph displays data as a histogram with bars that varied more from a normal curve (bumpier), with both a smaller range and with fewer distinct values; this histogram represented another alternate conception about the meaning of variability—one that is less correct from a statistical standpoint, but possibly closer to a colloquial interpretation of variability as being less predictable. To find out which meaning of "variability" they were considering when they chose their response, they were also asked to state their reasoning. On the pretest, results revealed that a little more than half of the prospective teachers in the class (61%) responded correctly at the beginning of the course by choosing the first histogram. Nearly all students chose the correct histogram on the posttest (89%), but few of these (2 of 16) stated that it was due to the data deviating more from the center. Most of those choosing the correct histogram replied it was due to the distribution having a greater range (10 of 16) and/or because it had a greater number of distinct values (6 of 16). This response reflects an awareness of variability as spread as opposed to "bumpiness", but may be less technically correct. The conceptual confusion of variability as "bumpiness" was pointed out during the class, which likely explains why so few students chose the second histogram on the posttest.

4.2.6 Question 23 (Hospital Problem)

Another problem that was particularly interesting was the Hospital problem (adapted from Kahneman, Slovic, & Tversky, 1982) which is commonly used to examine subjects understanding of the relationship between variation in a proportion and sample size (Garfield, 2003; Garfield & Gal, 1999; Rossman et al., 2001). The problem statement is given below:

23. 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)

Correct answer: B

A frequent response to this question is that since the probability of a boy being born is essentially equal to that of a girl, then either hospital is equally likely of having a day where 60% of births produce boys. This does not take into consideration, however, that the sampling distribution of proportions depends on the sample size. This question is situated in an everyday context yet contains very sophisticated underlying statistical concepts. Performance on this problem can be summarized in the table below:

Table 4.7: Number of students (percentage) in each response category on the Hospital Problem (Kahneman, Slovic, & Tversky, 1982).

Hospital Problem		Pretest		
		Correct	Incorrect	Overall
Posttest	Correct	5 (28%)	8 (44%)	13 (72%)
	Incorrect	0 (0%)	5 (28%)	5 (28%)
	Overall	5 (28%)	13 (72%)	18 (100%)

Table 4.7 indicates that a little over half of the class (10 of 18 students) showed no change on their performance, either answering the question correctly on both the pre- and posttest (5 of 18 students), or answering the question incorrectly on both (5 of 18 students). No one who answered the question correctly on the pretest missed it on the posttest, but most students (8 of 13 students) who missed it on the pretest went on to answer it correctly on the posttest. Because this problem is such a classic, and particularly because it assesses the concept of variation, it deserved a little further analysis. Since this is such a popular question in the literature on assessments that

assess statistical reasoning, I conjectured that perhaps it is a good predictor of statistical reasoning. In fact, there was an interesting relationship between performance on this question and that of the other questions on the posttest. Those students who answered the Hospital problem correctly on the posttest (n = 13, $\bar{x}$ = 17.9, s = 2.09) had significantly higher mean score on the rest of the posttest (p = 0.02, Figure 4.7) than those who answered the question incorrectly on the posttest (n = 5, $\bar{x}$ = 15.4, s = 1.48). This implies that it may be a good assessment item as an indicator of an underlying understanding about variation and may be the reason other researchers frequently use this question in assessments.

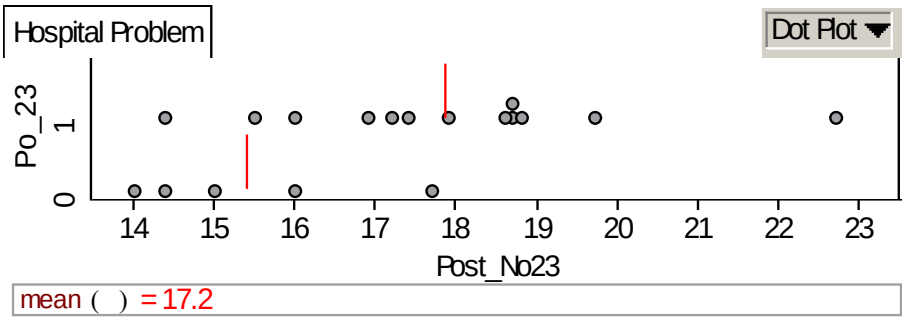


Figure 4.7: Performance on posttest (with Question 23 omitted) for those who answered Question 23 correctly verses those who did not.

4.2.7 Questions 11 (Dice Problem) and 25 (New Zealand Problem)

Finally, an interesting comparison can be made between two problems, Questions 11 and 25, that appeared in Pfannkuch and Brown (1996). While these problems are structurally equivalent, they are set in different contexts:

11. A six-sided die is thrown 7 times resulting in the following outcome: 3, 3, 3, 4, 4, 5, 5 (order is not important). Do you think there is evidence to suspect that the die is unfair? Why or why not?

25. Every year in New Zealand approximately seven children are born with a limb missing. Last year the children born with this abnormality were located in New Zealand as shown on the map below. Note that the population in each region below is approximately equal. A group of families in the central regions have filed a legal case claiming the incidence in their region is unusually high. Do the data support their claim? Why or why not?

Pfannkuch and Brown noted that even though these problems draw on the same statistical knowledge, the context of dice encourages more tolerance in variation for small subgroups, while less variability is tolerated for real-world contexts. Similar results were found in this study. On the pretest, 67% of students responded correctly on Question 11, indicating that they would need more rolls to be able to determine whether or not the die was fair. Only 17% of students answered correctly on Question 25, however. Of the 12 students who answered Question 11 correctly, only two (17%) also answered question 25 correctly. The gap narrowed on the posttest, with 83% of students answering Question 11 correctly and 50% answering question 25 correctly. This indicates that, as Pfannkuch and Brown found, the context of the situation in the question can affect whether or not people are able to think probabilistically, particularly novices.

4.3 SUMMARY

The results presented here indicate that the improvement in understanding of concepts of variation and distribution, as measured by the pre-post tests, was significant. The most striking finding was in the results of levels of confidence by minority students before and after the course. The experience in the course, heavily emphasizing concepts of equity and inquiry, appears to help minority students develop a level of confidence in their understanding of statistics closer to their Caucasian classmates and may have erased significantly lower levels of confidence recorded at the beginning of the course. This is critically important when one considers the literature in

equity that indicates confidence is critical to students' decision to enroll in higher level math and science courses (Lynch, 2000, citing Eccles, 1995). However, the results here are only exploratory and need further research to confirm whether this sort of experience combining equity, inquiry, and data analysis can assist minority student confidence in statistical understanding.

Also encouraging on the results of the pre-post tests was growth in understanding of deeply contextual problems, which consistently posted among the lowest performances on the pretest. This is a critical area of growth for general statistical literacy in applying concepts of variation in everyday contexts. The finding here in comparing the same statistical concept in a dice context and a media context was consistent with the finding of Pfannkuch and Brown (1996) that students use different intuitions about variation in these two contexts. Most students are comfortable thinking probabilistically in dice and coin problems, with which they have personal experience with variation, but tend to think deterministically if the same problem is posed in a real-world context one might find in the media. Finally, the consistent growth in questions that involved properties of skewed distribution and measures of center in histograms may point to a problem in some applied statistics courses of relying too heavily on normal distributions so that students struggle to transfer basic concepts when the distributions are either not symmetric or not theoretical.

The next chapter will examine the qualitative results from data collected during interviews and the final inquiry project to further probe understanding of variation outside of a closed, pencil-and-paper context. Furthermore, the results presented in this chapter will be used to corroborate the value of the qualitative results.

Chapter 5: Qualitative Results

Chapter 4 presented results that examined the prospective teachers' understanding of statistical concepts of variation and distribution in the limited medium of the pre-post test, focusing on the first part of the first research sub-question (below). This chapter seeks to use qualitative data to better understand the remaining sub-questions posed in Chapter 3:

1. How did the teachers express their understanding of concepts of distribution and variation?
2. What is the potential for technology in enabling teachers to conduct an inquiry in a semi-structured environment? What behaviors did the teachers exhibit in using the technology?
3. What can be said about preservice teachers' understanding of equity from their structured and ill-structured inquiry activities?
4. What is the interplay between the preservice teachers' statistical reasoning and the depth of self-designed inquiry into complex, ill-structured problems?

First, I will show the results of interviews conducted before and after the course that will document how the teachers articulated their understanding of variation and distribution as they compared two groups (Section 5.1). These results have also been published elsewhere (Makar & Confrey, 2003; under review) in response to a call for research focused on reasoning about variation (Garfield, Ben-Zvi, & Mickelson, 2002). Second, I will report results dealing with the behaviors teachers engaged in while using the Fathom software during a semi-structured investigation (Section 5.2). These results come from an interview conducted at the end of the course. This will give some insight into the ways that the prospective teachers engaged with the software. Finally, the results of the final projects and presentations will be reported to show how the teachers expressed their beliefs about equity, concerns and conflicts they encountered in dealing with the topic of equity, what they learned about equity during the inquiry process, their engagement with their inquiry topic, and examples of how they used concepts of variation and distribution to express their understanding of equity (Section 5.3). The chapter will conclude with an examination of interactions between the teachers' level of statistical evidence, use of the software in their semi-structured investigation, and engagement with their inquiry project.

5.1 ARTICULATING VARIATION

What did the prospective teachers pay attention to when they compared distributions? How did they articulate their understanding of distribution and variation when comparing groups? Although the pre-post test was designed as a measure of their understanding in this area, it did not probe into *how* that the preservice teachers would articulate distribution or variation, that is, the language they would use to talk about distribution and variation. This section strives to document the teachers' formal and *informal* notions of statistics, and specifically of variation and distribution—in particular, to address the second part of research sub-question #1: What concepts of distribution and variation did the teachers learn, and how did they express their understanding?

In this section, I will present how the prospective secondary mathematics and science teachers use their own words to compare the performance of two groups of students on the TAAS exam. From interviews conducted at the beginning and end of the semester, their descriptions of the data distributions under investigation are analyzed and categorized. Their language is often informal, but reflects a strong emergent intuition about variation and distribution in data. In this section, after introducing the interview task, I will document first their descriptions based on more standard, formal statistical language—mean, percent or number improved, outliers, distribution shape, standard deviation and range—then report their descriptions using more informal, non-standard statistical language. These descriptions fell into three categories, what I call *clumps*, *chunks*, and *spread*.

The purpose in reporting the results in this subsection is two-fold: (1) to document the various ways in which the teachers expressed "seeing variation" (p. 10, Watkins, Schaeffer, & Cobb, 2003); and (2) as evidence, along with the pre-post test, that as a group, the prospective teachers articulated a rich, although perhaps informal, understanding of distribution and variation when describing graphs and comparing groups. These results will be contrasted with the teachers' inclusion (or lack of inclusion) of variation and distribution to support their findings in their final inquiry projects (Section 5.3).

5.1.1 Interview Task and Analysis

The task given to teachers was set in the context of a school trying to determine the effectiveness of a semester-long remediation program (called 'math enrichment') for eighth-grade (14-year old) students thought to need additional help preparing for the state exam given each spring. During the interview, the prospective teachers were provided the following description of the situation (see Appendix G: Post-interview Questions) and shown Figure 5.1 (below):

> A local urban middle school has created a program to assist students who need extra help to prepare for the TAAS Math exam. They meet as a regularly scheduled class called "Math Enrichment". The students were placed in the class if their counselor determined they needed it. The school is interested in whether the program is helping students to improve their scores and have collected data on the difference between their 7th grade TAAS math test and a practice TAAS test given to them in the spring of the 8th grade. A graph of the change in scores is shown below. A positive difference indicates that the student scored BETTER on the 8th grade practice math test than on the 7th grade math TAAS. The data from the students in the enrichment class are on the top distribution and the data from the students not in the enrichment class are on the bottom distribution. The mean improvement of each group is shown on the graph. In addition, students highlighted in red are classified as Economically Disadvantaged.

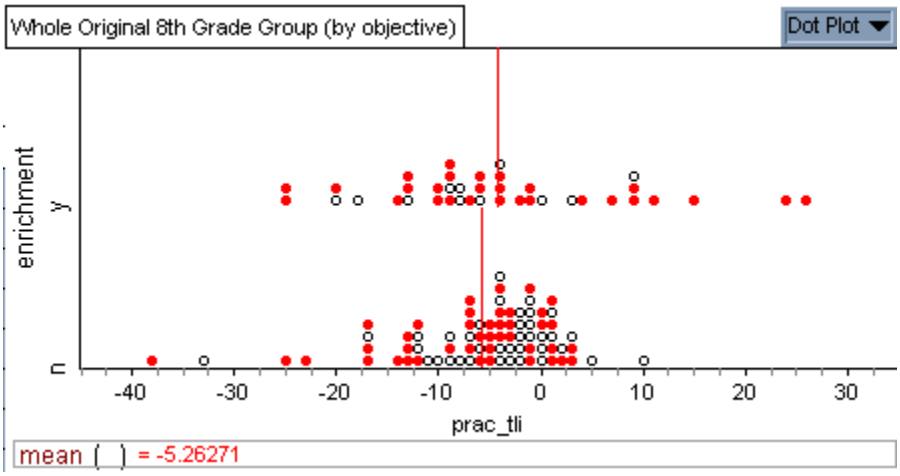


Figure 5.1: Graph shown to subjects during the Fathom interview task.

Next, the subjects were prompted to "Compare the improvement of the students who were in the enrichment class with those who were not." Since the same task was given at the beginning and end of the course, it is possible that any changes in the teachers' descriptions between the January interviews and the May interviews are due to a practice effect, however, many of the prospective teachers indicated that they did not remember the task.

Transcripts from the interviews were subjected to line-by-line open coding using NVivo (QSR, 1999) to seek the types of descriptions the prospective teachers used to compare the distributions. Thirty initial codes were created from this analysis which depicted the kinds of descriptions that emerged from the subjects' transcripts. These codes were further subjected to axial coding to attempt to seek relationships between these coded statements to create major categories of descriptions. Codes fell into two major categories that were identified in the kinds of statistical terminology, standard and non-standard, that the subjects used in their descriptions. Dimensions of these categories were further sought through selective coding to look for examples and structural variations within these major categories.

In comparing distributions in this context, anticipated results were comparison of means, measures of spread, proportion of students who improved, or descriptions of shape. All of these standard descriptions were found among the transcripts. However, more commonly, non-standard descriptions were given. These were more difficult to interpret because the terms used in these descriptions are not well-defined (Reading & Shaughnessy, in press), although they clearly held meaning to the teachers who used them.

5.1.2 Standard Terminology

With regard to conventional terminology, I will provide examples of where subjects used traditional concepts with formal definitions in statistics to compare the relative improvement of each group. For example, descriptions that include means, range, standard deviation, percentage improvement, shape, or outliers.

Means

Only a little more than half (12 of 21) of the prospective teachers interviewed used means in their descriptions in January, despite the fact that the means were marked on the figure for each group and additionally pointed out by the interviewer when

describing the task. This was somewhat surprising when you consider that 12 of these 21 teachers (and only 7 of the 12 who mentioned means) had already had either a formal statistics course or some treatment of statistical methods within a math or science course, and thus should have been trained to observe means.

Use of the mean in the comparisons for these eleven teachers ranged from a brief mention to a major focus of the discussion. For example, some of those who made use of the mean, did not use any other statistical descriptions to compare the distributions in January:

> *Mark*: Well, it looks to me like, uh, the group that did the Enrichment program overall has a better, uh, improvement even though it's not really even- [an improvement].
>
> *KM*: Okay. … And what are you basing that on?
>
> *Mark*: Uh, cause you. I think you said that this line was the mean? … So, uh, I was looking at that.
>
> *José*: It seems about even. I mean, they didn't decrease by that much, compared to the other [group]. … I don't even know what that would be, a point between their mean and their mean?

Both Mark and José had previously taken a formal statistics course, and José was one of two prospective teachers in the study whose analysis focused solely on means.

In previous work (Makar & Confrey, in press), it was found that although some teachers had a deterministic view of measures, others indicated some tolerance for variability in means, as did two teachers who recognized the effect a small sample could have on the mean. Note that the first excerpt below, from Angela, comes from a teacher with no formal training in statistics whereas the second teacher, Janet, had taken a traditional statistics course previously.

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

It is not clear from Angela's interview whether she was indicating that the mean might have been different if the sample was larger because of a recognition of variation in samples, because of the proportional effect of outliers, or another perspective. In hindsight, it is unfortunate that the interviewer did not ask her to clarify this distinction. Janet put forward a similar view that sample size might affect the mean, but articulated her reasoning more clearly:

> *Janet*: So the Enrichment class did have a higher mean improvement, higher average improvement, uh, but they had a smaller class. Um, I don't know what else you want me to tell you about it.
>
> *KM:* You said they had a smaller class, is that going to have any-
>
> *Janet*: A smaller sample size can throw things off.
>
> *KM:* How's that?
>
> *Janet*: (laughs) Um. The. With a, a larger population the outliers have less of an effect on the, on the means than in a smaller sample. So it doesn't, um, I don't remember how to say it, it doesn't, uh, even things out as much.

Janet's initial statement "I don't what else you want me to tell you about it" suggests the possibility that she saw a difference in the means, but little else worth discussing. Her observation that the outlier will have a greater effect in a small sample may indicate a perception that the variability of measures is greater in small samples. Another possibility, which could coexist with the above perception of variability of measures, is that her concept of mean and sampling is grounded in proportional reasoning. That is, the outlier makes up a larger proportion of a small sample than it would in a larger sample, and thus has a greater effect on the value of the mean.

Kathleen, who had also used some statistics before in science, recalled comparing means there:

> *Kathleen*: The mean [Enrichment] was a little bit higher than the, the group who didn't, who didn't take the Enrichment class. And I don't know if that would be statistically higher, but-
>
> *KM:* What do you mean, 'statistically higher'?
>
> *Kathleen*: Like if you, if you ran statistics on it. Like a t-test or something.

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

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

Although through further probing she was unable to articulate what she meant by 'normalize the data', it seems likely that Kathleen was referring to the dependence of sample size on key outcomes of the Central Limit Theorem to compare means with sampling distributions.

In the May interviews at the end of the course, fifteen of the eighteen teachers interviewed made some mention of the means, often with more specificity.

> *Anne*: Well, it looks like the students in the Enrichment class, on average, um, improved, or didn't decline as much as the ones in the regular class. Um.
>
> *KM:* And what are you basing that on?
>
> *Anne*: The means. Uh, the regular class is down by negative, uh, seven, six, minus six. And the Enrichment on average is at minus, um, is that three?

Percentage or number improved

Overwhelmingly the most common (15 of 21 teachers) comparison the teachers made in January was of the percentage or number of students in each group whose scores improved or dropped; two teachers' January analyses focused solely on reporting improvement. In many cases, teachers split the groups into two—improved or not improved—as exemplified by Rachel and Hope:

> *Rachel*: Well, it looks like more percentage of the people in the Enrichment program improved. Well, not overall, but like there were, if you look at the percentage. … It looks about half and half, or maybe a little bit more did not improve in the Enrichment program, but it looks like the ratio of the improvement is higher in the Enrichment program.
>
> *Hope*: Some people have improved. A lot of people.

In May, the use of the criterion of proportion of students improving persisted in the prospective teachers' comparisons of distributions, although many of the teachers were more likely to quantify their descriptions and none of the preservice teachers relied on proportions as their sole piece of statistical evidence.

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

Although mean and percentage improvement are important considerations when determining the effectiveness of a class targeted to help students improve their test scores, my hope was that teachers would do more than just reduce the data and compare means or percentage improvement as their sole method in determining how well the Enrichment program may have worked, something that four of the teachers did in the January interviews. Instead, I sought a more robust understanding of statistics within the context and an examination of the whole distribution in describing their comparisons in this context. For example, since the two groups of students in the data were being compared based on their improvement from their 7^{th} grade test score to their 8^{th} grade *practice* test, they might have explained the negative mean improvement of the students by recalling a discussion from the course that students might not take practice tests as seriously, knowing that their score does not have significant impact on them.

Outliers

After the category of percentage or number of students who improved, the next most common category used in January by the prospective teachers to describe their comparisons of the two groups arose through examination of outliers. For example,

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

> *Kathleen*: Well the group of kids that did take the, the [Enrichment] class, had an overall, I guess, didn't have the outliers down in the negative forties and, … who, I, I guess you could make an assumption and say, were motivated. That would just be an assumption, but, say that they were motivated enough to, um, to do well, and were all the way up here on the ten, twenties, and thirties [in improvement] on the test.

> *Anne*: It looks like a few students responded really well to the enrichment class and improved their scores a lot. … [And] for this, these, this student in particular, but all of these [pointing to outliers on the right], the enrichment program worked. I would say that.

All three of these prospective teachers had previously had some statistics. Although Andre and Kathleen included both upper and lower outliers in their descriptions, it was more common for teachers to mention just one set, as in the case of

Anne. The descriptions by other teachers in the course with no statistics background were less precise as exemplified by April, Gabriela, and Carmen:

> *April*: Well, it seems like the kids that needed the extra help a lot more [i.e. Enrichment] did a lot better.
>
> *Gabriela*: There's only, like these two out here that have actually, like, greatly intensely improved, and these have improved too, but, it's just like the areas seems like none of them really improved.
>
> *Carmen*: Some people improved up to, almost up to thirty points and then, but then, you know, on the non-enrichment class nobody is up here, so I would say that, um, the Enrichment class definitely had a better performance.

These three teachers focused on not just the criterion of *whether* students improved, but qualified it with *by how much*, indicating that they were seeing the upper outliers of the distribution and not just whether or not the data fit the criterion of improvement. Note that the preservice teachers above are not focusing on individual points, but a set of points that satisfy a criterion of "doing a lot better" or having "greatly intensely improved". When focusing only on the number or percentage of students that improve, one does not consider how far these values of the data are from the zero, only how many are above this point. That is, a student who improves by only one point is counted no differently than one who improves by thirty points. I would argue that descriptions that included qualities of the distribution that indicate attention not just to whether students improved, but by how much, are a richer descriptions in the context of testing.

Shape

Although the shape of a distribution is a critical attribute, there was not much variety in the descriptions that emerged that involved standard descriptions of shape, so their summary here is brief. Standard terminology describing shape (e.g. bell-shaped, symmetric, skewed) were uncommon in January (2 of 21 teachers), and somewhat more frequent in May (7 of 18 teachers):

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

Standard Deviation

Only two of the teachers in the study made any mention of traditional measures of variation (e.g. interquartile range or standard deviation) in the interviews, both at the end of the course. Although it was introduced in class and included in a homework assignment (p. 93-109, Rossman, Chance, & Lock, 2001), this may be due to the difficulty of these concepts or because the design of the task did not necessitate the use of these measures. In one of case, a teacher mentioned standard deviation, not as part of any particular argument, but to state its relative size in each distribution:

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

The other teacher who mentioned standard deviation indicated that she knew the term, but implied that she did not see it as useful in comparing the relative improvement of the students in the enrichment program with those who were not, stating a few minutes into her interview:

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

Charmagne's rather procedural description here seems to indicate that she was including the shape and standard deviation because she thought I expected her to discuss these, and her laugh seemed to acknowledge that these statements did not hold meaning for her in this context.

Range

Ten different prospective teachers used the term "range" during the January and May interviews. In almost every case, their use of it was linked to notions of scale and location. Here, scale refers to a description relative to the scale of the distribution (e.g. above zero); location refers to a description that indicates the relative location of the points being considered (e.g. "in the middle").

> *Carmen*: I would say that, um, the Enrichment class definitely had a better performance since most people are concentrated in this area, whereas you have a wider range and even a very good amount of points that they improved on. … I think it is working, yeah. Because you have a just wider range, whereas everyone was kind of close in on their improvement here. Uh. With the wider

> range, um, I would say it's working at least for some of the students. Because in the non-Enrichment, no one seemed to improve that much. … Because there's not a range here. [January]

Carmen's use of the word "range" here gives a sense of relative location. First, she contrasted "wider range" with "concentrated in this area" and "close in … here" giving the impression that she was indicating that the data were spanning a greater part of the scale. Next, her suggestion that the wider range implied it was "working at least for some of the students" indicates that it was located in the upper part of the scale, unlike the data for the non-Enrichment group where she said "there's not a range here", meaning in the upper portion of the scale. Brian, in January, also used the term "range" to indicate the numerical scale:

> *Brian*: It seems pretty evenly distributed across the whole scoring range.

In the May interviews, the use of "range" was more common, even though it wasn't a term used formally in the class. In almost every case, the term "range" either meant an interval (as in the case of Carmen and Brian above) or the length of an interval.

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

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

One difficulty may be that in school mathematics, the term "range" of a function is defined as a set, almost always an interval on the real number line. In statistics, however, the term "range" is a measure—the absolute difference between the minimum value of a distribution and the maximum value. By using the same term to indicate a set and a measure, one can begin to see where the distinction between objects and measures become murky in statistics. In school, the distinction between a geometric object (like a polygon), a measure of it (its area or perimeter), and a non-numerical attribute or categorization of it (closed or convex) is made clear. In teaching statistics, there is often not a clear distinction between an object (e.g. a distribution), a measure of it (its mean

or interquartile range), and an attribute (e.g. its shape) among less formal concepts of variation.

The results above are not meant to downplay the importance of teachers' use of standard measures in comparing groups. On the contrary, these are very important tools. The hope was that teachers do not emphasize simply summarizing or reducing the data to make comparisons as has been seen in schools when examining test data (Confrey & Makar, in press), but rather seek insights into the context the data represent through richer views that include notions of distribution and variation. This is particularly true given the evidence that students often use standard statistical procedures without considering whether they are appropriate (Gardner & Hudson, 1999; Abelson, 1995) and may lean towards using standard vocabulary even if it doesn't hold meaning for them or is not necessarily useful for describing the problem at hand. This next section will begin to look at the prospective teachers' non-standard and informal use of language, as this may provide greater insight into how their individual conceptions of variation and distribution hold meaning for them (Lemke, 1990).

5.1.3 Non-standard Terminology: Clumps, Chunks, and Spread

In this section I introduce three categories of descriptions the prospective teachers used in making their comparisons. These categories are more informal, and refer to qualitative attributes of variation and distribution rather than quantitative ones. The first of these categories is a *clump*. By clump, I mean an area of the distribution where the data is denser, usually (but not necessarily) the portion of the distribution surrounding the mode. The second category, a *chunk*, refers more generally refers to a contiguous subset, or partition of the whole of distribution. Therefore, all clumps are also chunks, but not conversely. The third category, *spread*, while much more commonly heard in statistics than clumps or chunks, was much more difficult to pin down. For some, the concept of spread may seem related to quantitative aspects of variation, like standard deviation or range. More commonly, however, the teachers used their own terminology to describe spread, capturing a qualitative aspect of variation and distribution that could not be separated from either. These three categories overlap both with themselves and with concepts of variation or distribution, cautioning that they should likely not be treated as distinct concepts.

Clumps

Although categorizing the data by percentage or number improved (described in Section 5.1.2) was very common, not all of the teachers used this criterion to split the distribution into two categories of improvement and non-improvement. Other teachers split each group into three categories: improving, not improving, and "about the same", as Chloe did in January:

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

Chloe's mention of "some on the same level" indicated that she saw more than just the students who scored above and below an absolute cut point of zero and that the demarcation between improving and not was more blurred. By seeing a middle group, she may have been seeing what Konold et al. (2002) call a *modal clump*:

> To summarize their data, students tended to use a 'modal clump,' a range of data in the heart of a distribution of values. These clumps appear to allow students to express simultaneously what is average and how variable the data are. Modal clumps may provide useful beginning points for explorations of more formal statistical ideas of center (p. 1).

Chloe's mention of a modal clump was not unique. For other teachers, notions of a modal clump also surfaced in January:

> *Janet*: [The Economically Disadvantaged students] seem pretty evenly distributed. I mean, from this bottom group here, they, the two kids with the highest scores are not Economically Disadvantaged, but that's just a, they don't have that much improvement above the others, above *the main group*.
>
> *Hope*: It's kind of flip-flopped *the big chunk* of this one is on this side, the big chunk of this one is on this side.
>
> *Chloe*: It seems like the most, *the bulk of them* are right at zero.
>
> *Gabriela*: Well, it seems like even though they were in Enrichment, like, there was still an improvement, or a lack thereof, like with the students that didn't take Enrichment anyways. … *The majority* of the ones that took Enrichment anyways are still more in the middle. [pause] Or they stayed the same, or they got worse, so I would say that just it's not an effective program.

The use of terms like *the main group*, *the big chunk*, *the bulk of them*, and *the majority*, by these teachers all indicate an awareness of a modal clump. In addition, Gabriela also noted that the distributions overlapped, by saying they were more in the middle, indicating that she was perhaps expecting a more distinct division between the two groups if, in fact, the Enrichment program was working. Gabriela seemed to be basing much of her tendency to split the distribution into three groups by focusing on the *scale* rather than on the notion of location: *low, middle, high*. In the May interview, this perspective changed as she added an additional caveat to her description:

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

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

> *Carmen*: All of these students improved a lot more than, uh, this big clump of students here in the non-Enrichment program.

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

The focus on number or percentage of students improving was very common in the teachers' comparisons (more common than reports of average). In addition, many of the teachers who compared the two groups based on the number or percentage that improved also saw a "majority" or "main group" clump. This may indicate that clumps are strong primitive notions of distribution. The fact that almost all of the teachers used some such criterion to separate each distribution (into either two or three groups) indicates that modal clumps may be a useful starting point to push understanding into a more distributional view.

Chunks

Next, I examine distribution *chunks*. Here the notion of chunk is informally defined to be a contiguous subset of the distribution created by a partition. As discussed above, some teachers split groups into two or three partitions. The clumps in the previous section, as well as outliers discussed previously, could be considered as distribution chunks.

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

Here, Maria partitioned the distribution into three categories, or "chunks": those who improved, those who "stayed about the same" and those whose scores decreased. Carmen, below, examines several different "chunks" in her description:

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

Carmen's descriptions in this early interview indicate she was not just paying attention to the criterion of whether or not the improvement was above zero, but also made note of the outliers "these two [that] improved so much" and those that "went way beyond the average". In examining *the majority*, and two different types of outliers, she indicates that she is changing her field of view from a fixed set of chunks to a more dynamic or fluid perspective. At her May interview, she continued this facility with moving between chunks, but was more specific in her description. A few of the chunks she mentions in this longer exchange are noted (Figure 5.2):

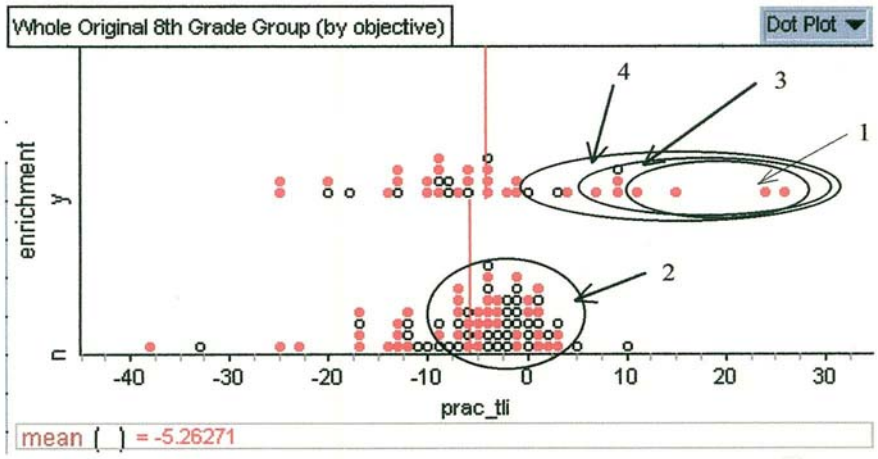


Figure 5.2: Carmen's description includes several subsets of the distribution, or "chunks", circled above.

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

Carmen: Um, well, … the student with the highest improvement in the enrichment program was about 16 points above the student with the highest improvement in the regular program. Um. And then, uh, this clump [upper Enrichment], there are others that are higher than the highest improvement, too. I mean there's about four that, um, improved more than the student in the regular program with the highest score [Figure 5.2, circled portion 1]. Um. It looks like, well there are more people, more students in the non-enrichment program, and, um, and the non-enrichment program, um, they also scored considerably lower. I guess the improvement was considerably lower than the students in the enrichment program. So. I don't know. …

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

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

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

Carmen: The same number. Around. Um. … I think, I think it is working and they should probably continue it because of the ones that improved, you know, well there's four that improved much greater than the, than the one that

> improved ten, by ten in the non-enrichment group, um, but then there were, there were about eight that improved, uh, more than majority of the non-enrichment group [circled portion 3]. So even though it wasn't, well, probably one-third, um, showed a considerable improvement [circled portion 4] and I, I would think that that's worth it.

Besides a quantification of her view of outliers (16 points above the maximum of the non-Enrichment group), one could argue that Carmen is seeing the set of outliers as more than just individual points, but as a contiguous subset of the distribution. As demonstrated above by the circled areas on the distribution, Carmen was also demonstrating her ability to see several *chunks*, with dynamic borders. This perspective of chunks (as a subset rather than individual points) is more distribution-oriented and indicates that examining chunks, beyond just the minimum and maximum, may be a useful way to encourage teachers to adopt a more distribution-oriented view of the data. Although again primitive, this notion of distribution chunks seems to fit somewhere between a focus on individual points and a holistic view of the distribution as a single entity or aggregate (see Konold, et al., 2003).

Spread

Of great interest in this study were ways in which the prospective teachers indicated notions of variation and distribution, particularly with respect to a particular context. But what counts as a notion of variation? One aspect of variation can be conveyed through descriptions of the spread of the data that might be captured quantitatively by the standard measures of standard deviation, range, or interquartile range. One might also conceive of variation as an attribute of a distribution (Bakker, 2004), like its shape. This section will explore the terms used by teachers in the interviews that capture their intention to communicate aspects of variation and distribution through a variety of words related to spread. The hope is that by comparing and contrasting their words, one can gain insight into how these non-statisticians describe and think about concepts of variation and distribution.

Andre and Margaret, both older college students with previous statistical experience, used the word "clustered" to describe their comparisons of the relative improvement of each group:

> *Andre*: I don't know what to make of this, actually, because as far as, like, it seems to me to support little difference between the Enrichment group and the

other group. Because. Um. Both groups are kind of clustered around the same area. [January]

Margaret: These are more clustered [pointing to the non-Enrichment group]. So where there's little improvement, at least it's consistent. This [Enrichment group] doesn't feel consistent. First impression.

KM: And you're basing this on?

Margaret: The clustering versus, it's like some students reacted really well to this, and some didn't. But it's more spread out than this grouping. Is, am I saying that okay? [January]

Andre's use of the term "clustered" highlights his observation that the *location* of the modal clump in the two distributions was the same—they overlapped. On the other hand, Margaret's initial description of "clustered" is paired with a notion of consistency, a concept Cobb (1999) argues is closely related to the notion of variation. She goes on to include it in contrast to being "more spread out", another term that is as difficult to define as "clustered". Five other teachers made use of the phrase "spread out" during the interviews.

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

Janet: They seem pretty evenly distributed, …fairly evenly spread out. [January]

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

These all use the word *evenly* with *spread out* implying that they may be indicating that the data were dispersed throughout the scale of the distribution equally, particularly given its common pairing with the phrase *evenly distributed* or *even distribution* in all three cases above. This context gives *spread out* a meaning related to the shape of the distribution, particularly given the contrast April made between *skewed left* and *evenly spread out*. Carmen's description below gives the phrase a similar meaning and José's, repeated below, is consistent with this interpretation:

Carmen: It's more spread out, the distribution in the Enrichment program, and they're really kind of clumped, um, in the non-Enrichment program. [May]

José: The standard deviation is going to be, like, really large on this, compared to that, because this is pretty spread out pretty far. [May]

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

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

KM: And you're basing that on?

Margaret: The clustering versus, it's like some students reacted really well to this, and some didn't. But it's more spread out than this grouping. Is, am I saying that okay? [January]

From the above six examples, the prospective teachers appear to be using phrases like *spread out*, *clustered* and *clumped*, as qualitative aspects of the shape of the distribution, similar to Konold's use of *modal clump*. I wondered whether the teachers' conception of spread, used above as an adjective, might be similar to their use of spread as a noun. Three cases were found of its use as a noun, by Carmen, Rachel, and Margaret, all in the May interviews:

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

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

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

Margaret: [describing the Enrichment distribution] It has a much wider spread or distribution than this group [non-Enrichment].

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

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

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

Hope's descriptions are akin to those heard above with phrases like *spread out* and *clustered* and *clumped*. Her use of the phrase *spanning all the way out to here* is unique (no one else used a similar phrase), but it appears that she is referring either to the range (in the sense of an interval) or making a point about the location of the outliers. Janet's and Anne's descriptions, like Hope's, also included the word *scattered*:

> *Janet*: There's Economically Disadvantaged kids pretty much scattered throughout both graphs. [May]
>
> *Anne*: I mean these are all kind of scattered out almost evenly. Whereas these are more bunched up together. [May]

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

> *June*: It seems like the, um, the disadvantaged [students] did a lot better scoring more towards this way than negative area and it seems that the people that weren't in the Enrichment seems to be all *gathered* from the zero and the negative side compared to the people that were in the Enrichment program because this is kind of *dispersed off* and this is like, *gathered* in the center. [January]

From these excerpts emerges a set of terms under the umbrella *spread* that indicate similar notions: spread out, scattered, evenly distributed, dispersed off. Antonyms that emerged include clumped, grouped, bunched up, clustered, gathered, tight.

5.1.4 Summary

This section documented the statistical language—both standard and non-standard—that the prospective teachers used in describing and comparing distributions. The evidence presented here, and from the pre-post test in Chapter 4, provides insight into the rich conceptions of variation and distribution that the teachers had or developed during the course. In both the pre-post test, and interviews documented in this section, the tasks were structured. Their conceptions here will be compared, in Section 5.4, to the prospective teachers' application of these concepts in the ill-structured context of their inquiry projects. I move now to the results of another semi-structured task to document the ways in which the prospective teachers made use of Fathom, the statistical software used during the course, to conduct a data investigation.

5.2 USE OF TECHNOLOGY

This section documents the results from a set of interviews conducted at the end of the course to probe into the teachers' behaviors when using Fathom to conduct a short investigation. As stated in Chapter 3, the Fathom investigations were designed to investigate how the prospective teachers would use dynamic statistical software in a structured investigation and addressed research sub-question 2: What is the potential for technology in enabling teachers to conduct an inquiry in a semi-structured environment? What behaviors did the teachers exhibit in using the technology?

The results in this section were obtained from a single task performed by the preservice teachers on the computer using Fathom during the final interviews at the end of the course. For the interview task, the author selected a random sample of 273 Hispanic tenth graders in Texas who took the 2002 TAAS exams and were living in either urban or rural districts. The data were authentic, with fourteen variables included, and sampled from a larger data set obtained from the Texas Education Agency; it contained raw and scaled scores for students on their 8^{th} and 10^{th} grade mathematics and reading tests, as well as demographic information. Before examining the data, the teachers were asked to make a conjecture about the relative performance of Hispanic students on TAAS in urban versus rural areas. After stating a conjecture, they were asked to go into Fathom and use the data (described to them briefly as above) to investigate their conjecture until they felt they had enough evidence that they could state a conclusion. The data file they used contained only a Fathom collection with the data; it was up to the teachers to make their own representations (e.g. graph or table) to examine the data. Most of the teachers created a dot plot similar to Figure 5.3 below.

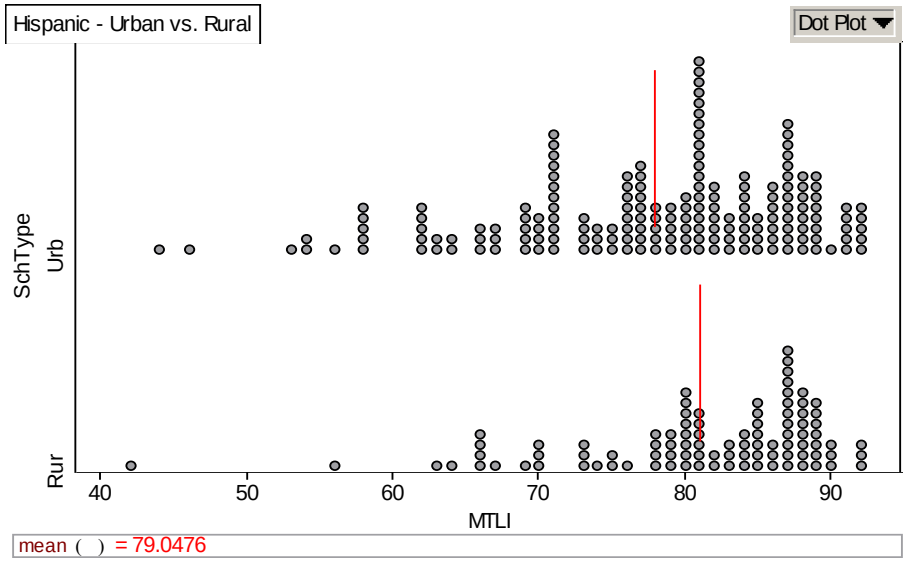


Figure 5.3: Dot plot of MTLI (scaled score on the mathematics TAAS test) split by School Type (Urban or Rural) for tenth grade Hispanic students in Texas. This graph is one similar to those created by several of the preservice teachers during a Fathom investigation.

Interviews were transcribed and subjected to line-by-line open coding using NVivo (QSR, 1999). Fourteen codes emerged from this analysis which captured their choices of evidence, representations created, types of investigations, and links between evidence and conjecture. This analysis led to the development of a theory about possible differences in the behaviors in the prospective teachers' uses of the technology in their investigations. Transcripts for each subject was then summarized, paying particular attention to conjectures, stated intentions, actions taken, representations created, feedback observed, conclusions that were drawn, and contextual explanations made. Some of these categories were influenced by Land and Hannafin's (1996; 1997) work on Open-Ended Learning Environments (OELE) which stresses attention to intention-action-feedback cycles to test conjectures created during an open-ended inquiry.

Three general behaviors types were discovered in the ways in which the teachers used the software to conduct their investigations, and the teachers were categorized by these behaviors types: the Wonderers, the Wanderers, and the Answerers. These categories were determined qualitatively based on the decision path each subject created from conjecture through evidence use to their conclusion. In general, Wonderers were lead through the investigation by their "I wonder" questions. They created a theory and used the data to test their theory. Frequently, their results also generated new theories and again they checked to see if the data provided evidence in support of these theories. The Wanderers, on the other hand, went through their investigation using the data to look for a theory. They spent a good portion of their time "wandering" through various analyses that were not necessarily directly connected to their conjecture, but hoping that something would jump out at them that they could tie back to their conjecture. The final group, the Answerers, went into the investigation with a theory, like the Wonderers, but did not generate "I wonder" questions during the investigation. They searched for a particular piece of evidence to support or refute their conjecture and then directly stated their conclusion (the "Answer").

Three dimensions of responses to the subjects' feedback from the software were chosen to re-code the transcripts of the investigations to try and capture the differences between the three behavior types quantitatively. Each code is identified below (Table 5.1).

Table 5.1: Description of codes used analyze Fathom interview transcripts.

Code	Description	Example
Observations	observational statements about results but not directly connected to the conjecture	April: "The sample size of the urbans is larger."
Evaluations	evaluative statements about results that related directly (using the context) to the conjecture	Carmen: "So there were more students not passing in the urban schools, by, uh, percentage, right?"
Conclusions	conclusions drawn, based on the results, to support or refute the teachers' conjecture	Gabriela: "It doesn't matter where the Hispanic students come from, an urban area or a rural area, they perform at about the same level."

Transcripts were coded by two independent researchers (the author of the dissertation and one other researcher) with 94.6% agreement. The two researchers then reached 98.3% agreement on a subset of 10% of the transcript statements that were discussed further, but did not try to resolve differences on the remaining 2232 statements. A summary of the author's results are given below (Table 5.2):

Table 5.2: Mean and standard deviation of the number of statements made for each code by the three behavior types.

Mean (sd) Number of Statements		Behavior Type			Overall
		Wonderers n = 5	Wanderers n = 8	Answerers n = 4	Mean (sd) n = 17
Code	Observation	16.2 (12.6)	10.5 (5.4)	4.8 (2.2)	10.8 (8.5)
	Evaluation	13.2 (6.4)	5.1 (3.6)	3.3 (1.7)	7.1 (5.8)
	Conclusion	6.0 (2.7)	4.3 (1.8)	2.5 (0.6)	4.4 (2.2)
Totals		35.4 (18.4)	19.9 (9.3)	10.5 (4.4)	22.2 (14.8)
Mean (sd) Time Spent (minutes)		28.4 (12.2)	12.7 (4.5)	5.7 (2.6)	15.6 (11.3)

In addition to the behavior types, which are discussed below, all of the prospective teachers demonstrated their ability to use the software successfully to conduct the investigation. Furthermore, all of the teachers were able to state a measurable conjecture, seek appropriate evidence to support their conjecture, and state a logical conclusion based on what they had found. Furthermore, none of the teachers expressed discomfort in the size of the data set, which had thousands of pieces of

information with fourteen variables (columns) and 273 cases (rows). This indicates that with a semi-structured task, the teachers showed facility with the software, understanding of the conjecture-evidence-conclusion process, and comfort with encountering large (clean) data sets. It is unknown how these results would extend to their ability to conduct an investigation in a less structured task where the teachers would also have to seek their own data set, choose an appropriate sample and relevant variables, clean the data (e.g. handle empty fields or reformat entries), wrestle with uncertainties and inconsistencies in their results, and interact with their own beliefs about their chosen topic. Some insight into their ability to conduct a less-structured investigation will be the topic of Section 5.3. I now turn to descriptions and evidence of the types of behaviors where the teachers displayed variation in their approach to using Fathom as a tool for investigation.

5.2.1 Wonderers

Wonderers are those prospective teachers who were guided during their investigation by the "I wonder" questions that emerged as they tested their conjecture. Their investigation was lead by theories that they developed; using these theories they went into the data set, using the result of their analysis to test and possibly revise or refine their theory (Figure 5.4). Their use of the technology was as a tool for inquiry, one that would support them in the process of testing and evaluating emerging theories.

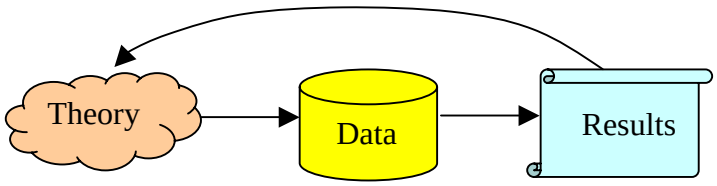


Figure 5.4: Model of Wonderer behavior.

Driven by desire to find evidence to support and refine their theory, Wonderers spent the longest time, on average, investigating their conjecture ($\bar{x}$ = 28.4 minutes, s = 12.2), but used their time purposefully. Because their investigation was goal-oriented, Wonderers had a significantly higher mean number of evaluative statements (p = 0.04) than Wanderers and Answerers, but not a significantly higher mean number of

observations (Table 5.2 and Figure 5.5), despite the fact that they spent an average of 18 minutes longer than other teachers testing their conjecture.

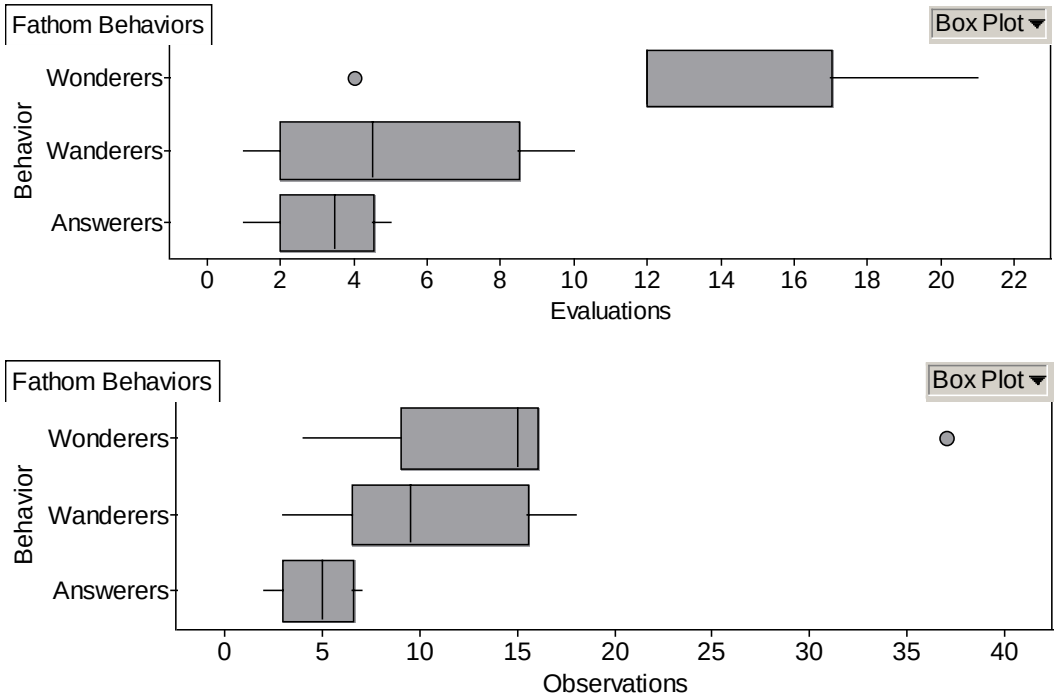


Figure 5.5: Box plots showing number of evaluative statements (above) and observations (below) for each behavior type: Wonders (n = 5), Wanderers (n = 8), and Answerers (n = 4).

5.2.2 Wanderers

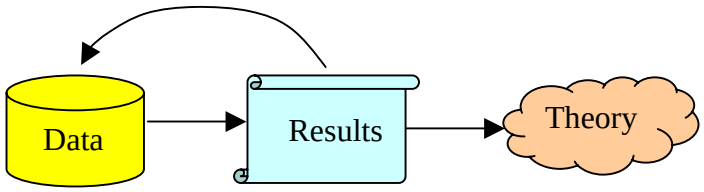
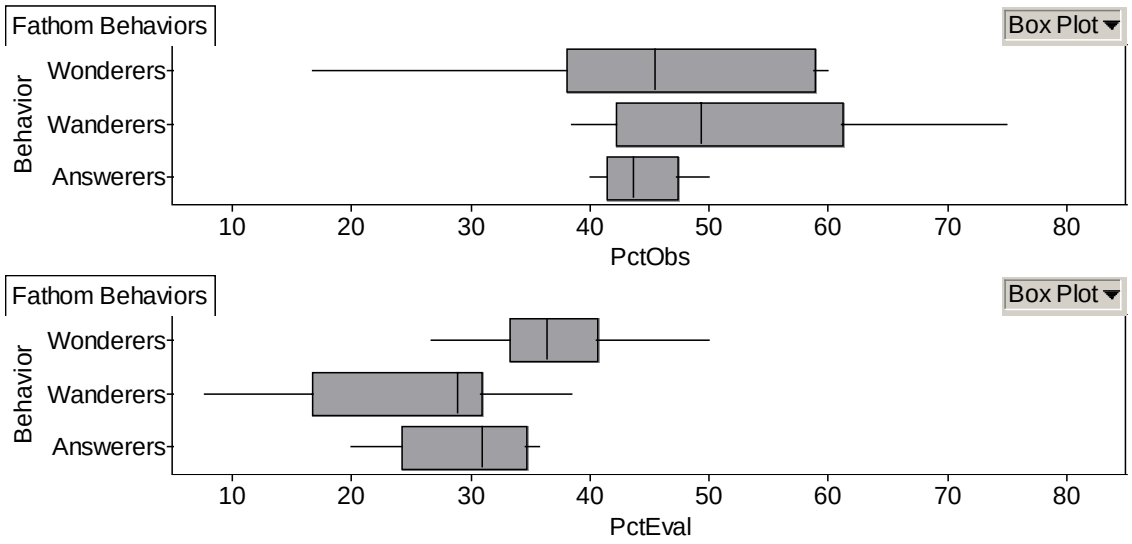


Figure 5.6: Model of Wanderer behavior.

Wanderers were identified by a tendency to look through the data to see if anything "popped out" at them, rather than going to the data with particular evidence they were looking for. Rachel, a Wanderer, summed it up well when she said during her investigation, "Well, I always like to look at everything" as did Christine who after

several minutes of investigating variables unrelated to her investigation, pondered, "Why am I looking at this?". The Wanderers' investigations included a conjecture, evidence, and a conclusion like those of the other behavior types, but their time was often spent wandering through the variables looking for patterns to emerge. Unlike Wonderers, who were driven by an internal theory, Wanderers were driven by the results that appeared from their wanderings. They went from the data set to graphs and then back to the data set until something "interesting" came up, leading them to develop a theory to explain the result. Their use of the technology was as a filter to "catch" potential sources of theory. In terms of measurable behaviors, Wanderers tended to fall between the Wonderers and Answerers in categories such as mean time spent and mean number of observations made. However, the mean percentage of their coded statements that were evaluations (Figure 5.7, bottom graph and table) was significantly lower ($p = 0.04$) than that of the Wonderers, even though the sample size was small.



Mean (sd) Percentage of Coded Statements that were:	Behavior Type			Overall n = 17
	Wonderers n = 5	Wanderers n = 8	Answerers n = 4	
Observations	43.8 (17.7)	52.4 (13.2)	44.3 (4.2)	47.9 (13.3)
Evaluations	37.4 (8.7)	24.9 (10.8)	29.4 (6.9)	29.6 (10.4)

Figure 5.7: Box plots and summary table showing the percentage of coded statements by each behavior type that were observations (top) and evaluations (bottom).

5.2.3 Answerers

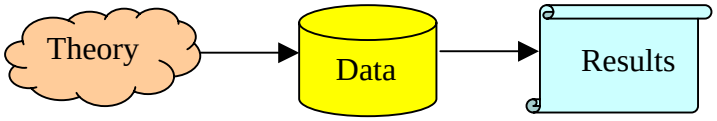
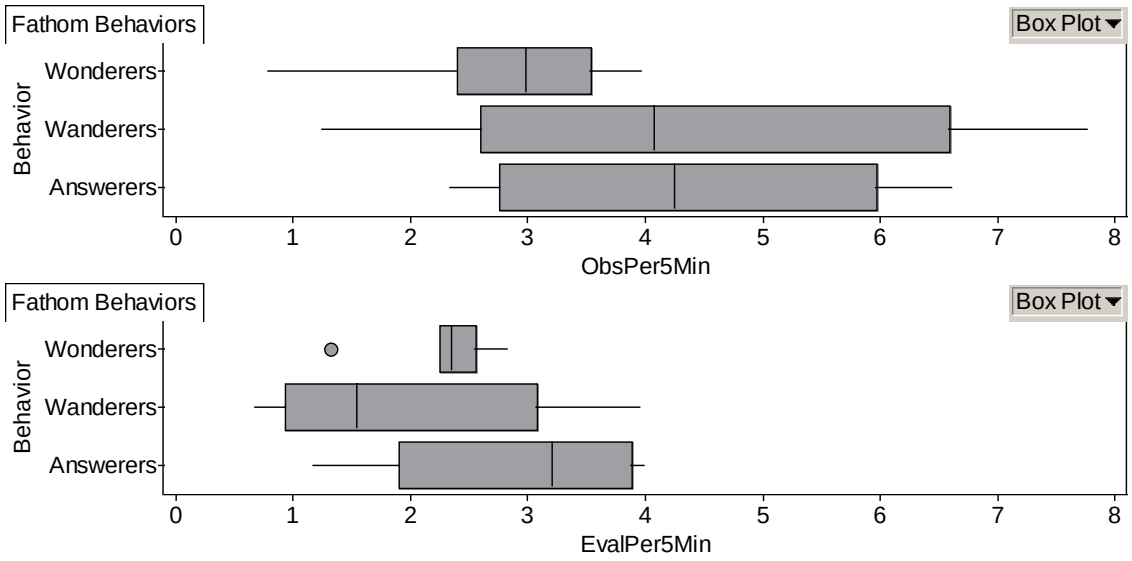


Figure 5.8: Model of Answerer behavior

The third group of behavior types recorded, the Answerers, used the software as a tool to locate a particular piece of evidence to test their conjecture and were then quickly ready to draw a conclusion. To this group, the computer was an efficiency tool that they could use to answer a question they had. This group was identified by their decision process: they looked for a particular, single piece of evidence and once they found it were satisfied that they had "answered" the question put to them.

As might be expected, Answerers clearly spent the least amount of time conducting their investigation, under six minutes on average—half that of Wanderers and a quarter of the time, on average, spent by Wonderers (Table 5.2). Even though they took much less time than Wanderers, they made a similar mean number of evaluations (Table 5.2 and Figure 5.5). Their *rate* of making evaluations was very different than the Wanderers (Figure 5.9), however, with Answerers making a median of 3.2 evaluations per five minutes, double that of the Wanderers, who made a median of 1.5 evaluations per five minutes. The difference in means was not significant ($p = 0.28$). Their median rate of making observations was very similar (Figure 5.9): 4.2 observations per five minutes for Answerers versus 4.1 for Wanderers.



Mean (sd), Median Number of Statements per Five Minutes that were:	Behavior Type			Overall n = 17
	Wonderers n = 5	Wanderers n = 8	Answerers n = 4	
Observations	2.73 (1.24), 2.98	4.44 (2.41), 4.08	4.36 (1.95), 4.24	3.92 (2.07), 3.18
Evaluations	2.26 (0.57), 2.36	1.96 (1.23), 1.54	2.89 (1.29), 3.21	2.27 (1.10), 2.36

Figure 5.9: Box plots and summary table showing the rate (number of statements per five minutes) of observations (top) and evaluations (bottom) made by each behavior type.

The results of a one-way ANOVA test showed that there was no systematic difference between the behavior types on their performance on the posttest ($F_{2,14}$ = 0.144, p = .87).

The Fathom interviews documented that all of the prospective teachers were able to create a reasonable conjecture, locate evidence in a given data set to support or refute their conjecture, and state a logical conclusion. Furthermore, while all of the teachers exhibited comfort with analyzing a rather large data set and facility with the software in conducting their analysis, the interviews uncovered a variety of approaches the teachers took in conducting their investigation. For some, the software was used as an efficiency tool to answer a question with evidence. For others, the software was used to connect with the context and motivate "I wonder" questions that served to allow the prospective teachers to dig deeper into the data for more subtle and sophisticated

relationships. The most common group, however, used the software without the same sense of purpose taken by the other two groups. In this final group, the affordances of the software may have been a distraction to the prospective teachers' investigations as they used the ease in creating graphs to wait for relationships to "pop out" of the data. It is possible, however, that without this software feature, these teachers would have struggled with even conducting a somewhat open-ended investigation that provided multiple avenues for analysis.

5.3 INQUIRY PROJECTS

The data collected from the inquiry projects were used to address the third and fourth research sub-questions:

- What can be said about preservice teachers' understanding of equity from their structured and ill-structured inquiry activities?
- What is the interplay between the preservice teachers' statistical reasoning and the depth and breadth of self-designed inquiry into complex, ill-structured problems?

In this section, I document how the pre-service teachers' investigations interacted with their understanding and inclusion of both equity and the statistical tools available to them. I conjectured that the depth of their inquiry would be strongly correlated with their understanding of issues of variation and distribution.

This section will first document the method of analysis of the inquiry projects. Next, the types of projects will be reported and the rubrics developed to describe level of statistical use and engagement with their inquiry. The inquiry projects will be reported in detail, including the beliefs about equity that emerged from the projects and reflection papers.

5.3.1 Method of Analysis and Organization

Extensive notes were taken on the transcripts of the videotape of each prospective teacher's class presentation of his or her inquiry results, written final papers, and interviews (for members of the focus group). These notes were subjected to line-by-line open coding (as in Sections 5.1 and 5.2 above) in NVivo. To maintain the focus of results on the third and fourth research sub-questions, the results of the open coding were first limited to issues related to beliefs about equity, personal engagement,

and statistical use, and then further subjected to axial and selective coding to investigate dimensions in these categories, develop rubrics, and locate examples. The rubrics were further used to describe personal engagement and statistical use (see below) and to aid in examining interactions with results of the posttest and technology behaviors.

Level of Statistical Use

The main research question was to look at how the preservice teachers used notions of distribution and variation to articulate their understanding of issues of equity. This section will examine the extent to which the teachers used statistics in their final project to present the evidence of their inquiry about equity. As the concepts of variation and distribution are at the heart of statistical reasoning (Wild & Pfannkuch, 1999), I should note that where the term statistics is used in these final two chapters it will be taken to mean the statistical concepts related to distribution and variation. The term "statistics" will be used for brevity.

A construct called "statistical level" can take on several interpretations. For example, "statistical level" in the context of inquiry could mean one of the following:

1. the sophistication of the design of their studies, including attention to data quality and method of choosing their sample, statement of assumptions and limitations of their study, and appropriateness of procedures used.
2. the quality of the chain of reasoning that the prospective teachers used in linking their statistical use to the problem under investigation. That is, one might rate whether there were clear links between the conjecture or problem they stated, the analysis they conducted, their interpretation of this analysis, and the conclusions they state.
3. the quality and depth of use of statistics as a means of reporting evidence for their findings. This is related to one's epistemology in his or her perception of the relationship between evidence and the acquisition of knowledge (King & Kitchener, 1994; Perry, 1968/1999). For example, the evidence they use might rely on personal experiences and beliefs, or on the authority and analysis conducted by "experts", or on their own analysis of the data. If they relied on their own analysis as evidence, one might rate the extent to which they engage the reader in the reasoning and interpretation of their analysis; for example, whether they report only the end result of their findings and let the data "speak for themselves" or take the reader through their reasoning.

4. the extent to which the prospective teachers used the concepts of variation and distribution in their analysis through the display and interpretation of the data in their analysis. This construct would rate, for example, whether the teachers compared groups by focusing only on comparing means of each group or if they chose to display and interpret the distributions of data for each group.

All four of these perspectives of the construct "statistical use" relate a mindset of data and statistical use that would be of value to develop in teachers. These four perspectives are likely related, but investigation of their relationship is beyond the scope of this study. In this dissertation, which is meant to explore the prospective teachers' use of the concepts of variation and distribution to support their inquiry into concepts of equity, the construct used will focus on the fourth interpretation of this construct: the perspective teachers' use of variation and distribution by examining the extent to which the prospective teachers used robust notions of these concepts in their analysis. This construct will be measured by examining the representations the teachers used in their final paper and the extent to which they included these concepts in their interpretation of the data. Five levels of this construct were chosen to record the teachers' statistical use in their final paper:

1. Examination of a single data point or single summary statistic, including maximum or minimum of a distribution;
2. Display of data or summary statistics in a table, or focus only on means or percent passing (even in a displayed distribution), or copy of a graphical image (e.g. from the Internet);
3. Use of a bar, circle, or line graph in a static display of summary data with no examination of trends or distribution;
4. Analysis using a distribution (dot plot, histogram, or box plot) or scatter plot including an examination of trends or variation;
5. Inferential statistics, for example with a simulation, including examination and discussion of variation

The choice of these levels is based on an assumption that examination of single data points is the least desirable activity and farthest from inclusion of variation and distribution. Evidence reported at the second level, including use of a tabular display or

reporting of summary statistics (proportions or measures of central tendency), is better than reliance on single data points, but numerical data often neglects concepts of variation and distribution (Confrey & Makar, in press). Copies of graphical images were also included in this level because they represent someone else's analysis and would not produce evidence that the teachers considered variation and distribution concepts themselves in their interpretation, unless explicitly stated otherwise. The third level, use of static displays, implies that the teachers appreciate visual representations in their ability to communicate the relative magnitudes of values. However, these displays (circle graph, bar graph, line graph) are only visual reproductions of the numerical information at level 2 and do not include rich notions of variation or distribution. The fourth level is the first one where teachers explicitly include the notions of these concepts in choosing to display, and include in their interpretation, representations that show the distribution and variation of data. Finally, the use of variation and distribution in a simulation implies that the teachers value that variation exists not only at the level of the data, but also for measures. This concept involves the notion of sampling distributions, a concept involving both variation and distribution that is very difficult to understand (Confrey & Makar, in press; delMas, Garfield, & Chance, 1999; Thompson, 2001; Saldanha & Thompson, 2001).

Projects were assigned their "statistical level" based on the following criteria: Each representation from their final papers was scored using the above rubric. The mean of the median level and level at the third quartile were used to determine the level of the project. The reason for this measure is twofold: (1) I did not want to discount a project's level if a subject provided *additional* tables or charts as evidence. (2) By choosing a statistical level above the median level of statistical evidence they used, I assume that a subject will not include (and discuss) higher levels of evidence than they see as needed to present their results. However, it is likely that subjects will include representations that score below their "optimal" level. (3) Because I would be comparing statistical level to other indicators (for example, engagement, pre-post test results), and the overall level of statistical use was skewed low, I wanted to maximize the spread of the distribution of levels of statistical use to increase the potential power of this measure.

The table below (5.3) indicates the number of subjects at each level of statistical evidence, rounded for simplicity. Note that in three projects, all at level 3, there were

two prospective teachers working together. In addition, one teacher (with a family crisis) did not turn in her final project at the end of the course although she did give a presentation, but there was not enough evidence to determine the level of statistics that she would have used had she turned in a written paper.

Table 5.3: Count and primary level of statistical use on inquiry projects by subjects

Level	Brief Description	Count	Percentage[3]
1	Single point or value	0	0
2	Data table or summary statistics	3	18%
3	Static display: Bar, circle, or non-trend line graph	9	53%
4	Univariate or bivariate distribution	3	18%
5	Simulation	2	12%

Admittedly, I was disappointed by the number of the prospective teachers who relied on static displays (level 3) or only summary statistics (level 2) even though they indicated on the posttest (Chapter 4) and interviews (Section 5.1) a fairly strong understanding and awareness of distribution and variation. Because more than half of the prospective teachers in the course had already studied statistics, I would have expected that they would have understood statistics at a more robust level than use of simple tables and graphs would indicate.

One explanation for why such a large portion of the class relied on only elementary statistical tools might be that the final project was a large portion of their grade (40%) and they did not feel comfortable taking risks by perhaps using statistical analyses incorrectly. Another possibility is that they were so uncomfortable with the process of inquiry, particularly with not being able to "answer" their questions, and the topic of equity, which they found very difficult to discuss, that these additional pressures distracted their energies away from concentrating on statistical evidence, reducing their analysis to elementary reporting of means or percentages. A third possibility is that they were *able* to use and understand a more sophisticated level of statistical analysis, but did not see it as a useful tool for inquiry. Finally, it is possible that they needed much more time and support than was provided during their projects in order to make the transition from in-class and short open investigations into a more complex and open-ended inquiry project. It was clear that many of the teachers spent a

[3] Note that these values may not sum to 100% due to rounding.

lot of their project time looking for data, settling on a topic, and developing their conjectures. In addition, some teachers procrastinated on their inquiry projects and then tried to write the entire paper in the final few days and found the process much more difficult than they expected. In the future, I would try to start this part of the process much earlier in the course, with greater accountability to stay on a more structured timeline, ensuring they had more time and more support to wrestle with the evidence, analysis, and writing portions of their projects.

From the table (Table 5.3) it can be seen that the majority of teachers relied on simple static displays that displayed only summary data. One might make the argument that the level of statistical use the prospective teachers used in their final inquiry projects was probably related to their understanding of statistics. Figure 5.10 below presents a scatter plot of performance by the preservice teachers on the pretest (left) and posttest (right) compared to the level of their statistical use on their final inquiry projects. The correlation coefficient indicates that the association between the statistical use and pretest is at best weak (r = 0.25) with little or no measurable association between statistical level demonstrated by the inquiry project and performance on the posttest (r = 0.04) or improvement from pretest to posttest (r = -0.20). This indicates that the level of statistical understanding, above a basic level as learned in the course, was likely not a factor overall in the prospective teachers' choice of evidence used in their inquiry projects.

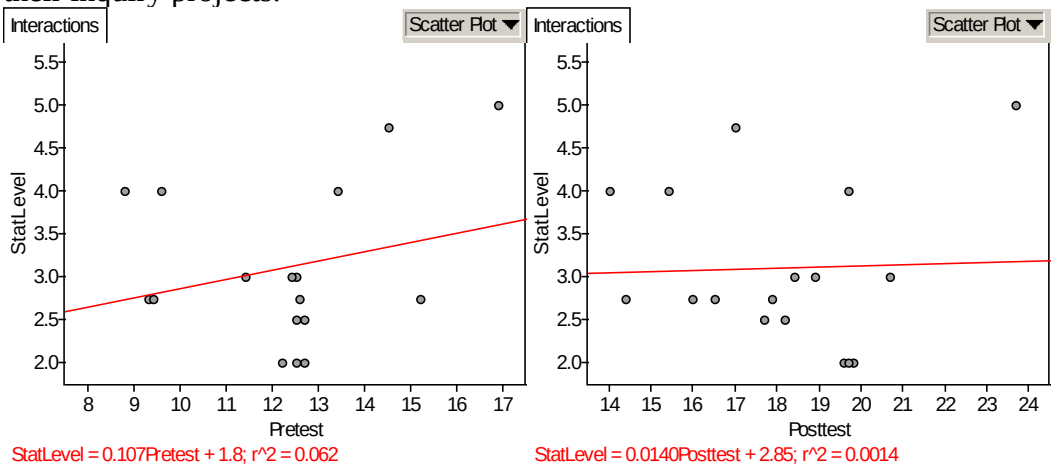


Figure 5.10: Association between the level of statistical use on the inquiry project and performance on the pretest (left) and posttest (right). The regression line is included for ease of visualization.

If it is assumed that the pre- and posttest are valid measures of understanding of variation and distribution, then the lack of correlation between statistical understanding and statistical use on the inquiry might be interpreted one of four ways. Either (1) the task motivating the inquiry project did not provoke teachers to use the level of statistics that they knew, (2) the experience of a complex inquiry of an ill-structured problem was so unfamiliar to them, or the topic of equity so uncomfortable, that it distracted them from the use of more sophisticated techniques, (3) understanding statistics does not imply that one feels compelled to use it as evidence, or (4) statistical understanding in a well-structured context does not necessarily translate into its use in an ill-structured context.

In comparing the behavior patterns displayed during the Fathom interviews (Section 5.3), it was speculated that there may be some differences in level of statistical use among those showing different behaviors as they conducted their investigation in Fathom. This relationship is represented in Figure 5.11. The graph shows that the Wonderers (n = 5) posted a higher average level of statistical use (median = 4) than the other two behavior types (n = 13, median = 2.75). However, under the null assumption that there is no difference in statistical levels used for the three Fathom behaviors documented, a one-way ANOVA indicates that there is insufficient evidence to reject the null hypothesis ($F_{2,14} = 1.57$, $p = 0.24$).

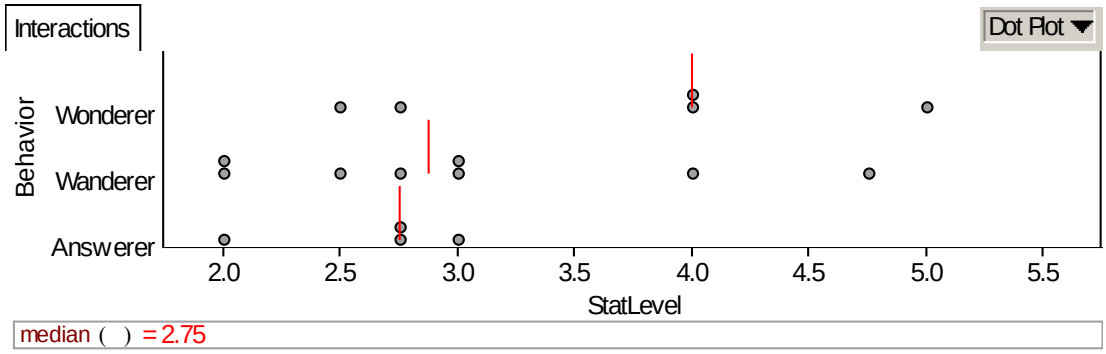


Figure 5.11: Dot plot comparing the level of statistical use on the inquiry projects by behavior type demonstrated in the Fathom interviews (Section 5.2). The median level of statistical use is marked for each behavior type.

Up to this point in the dissertation, the majority of the discussion has focused on the teachers' understanding and use of variation and distribution. However, the context

plays a particularly critical role in the inquiry projects that the teachers conducted. In order to better understand the inquiry projects in relation to their use of statistics *in the context of equity*. The next subsection will describe a scale devised to measure their personal engagement with their project investigating issues of equity.

Engagement

The task defining the inquiry project provided the prospective teachers an opportunity to choose a topic of interest to them within the bounds of equity and data-based inquiry. There are no assurances that being able to choose their own topic necessitated engagement in the topic. The construct of "engagement" will be examined as part of the descriptions of the inquiry projects and will consist of evidence of the prospective teachers' level of engagement with the inquiry task through either a personal connection to the topic (such as relationship of their topic of inquiry to their personal identity or beliefs, or to their life at home, work, or school) or unusually high interest in conducting the inquiry.

Nearly all of the prospective teachers chose a topic for which they found a personal connection, with few exceptions. They expressed their connections to their topic of study either in their presentation, or in the introduction to their written final paper, where they were asked to describe why they chose their study. The personal connections they conveyed varied from experience with the schools they investigated, as in the case studies conducted jointly by Angela and Gabriela, and by Sarah and April; connections to the school type under investigation, such the inquiry conducted by Brian who attended a magnet school and wanted to compare them to non-magnet schools; Rachel's investigation of the Robin Hood plan, legislation which her father had investigated for their local community council; Kathleen's investigation of dropout trends motivated by concern about a close high school friend who had dropped out of school; Janet and Margaret, each expressing an interest in teaching in an urban district and wanting to investigate issues they had heard about in urban districts and compare them with suburban districts, with which they had greater familiarity; or personal issues of race and class such as the inquiries conducted by three minority women: Chloe, Maria, and Charmagne. The projects will be discussed below. For a few teachers, it was unclear to the researcher what the personal connection might have been as the teachers did not make it explicit. Three inquiry projects in particular stand out: José's comparison of funding for schools rated low-performing and exemplary, Anne's

investigation into treatments undertaken by low-performing schools, and an examination of factors influencing Hispanic performance conducted by Emily and Mark, both from middle class White families.

The construct of engagement was measured holistically on an ordinal scale based on two elements: the closeness of the teachers' personal connection to the topic and their observed level of enthusiasm towards the task during the process of inquiry. The first of these elements, personal connection, was based on the level of the teachers' stated personal connection to the topic of study (during the inquiry process, in their presentation, or on their final paper), that is, how close the topic of their investigation was to them personally. The level of their personal connection was gleaned from the final paper's introduction where it was required that the prospective teachers describe how they chose their topic of inquiry. From this, as well as comments gleaned from their presentations, the teachers indicated either no particular personal connection, moderate connection, or strong personal connection to their topic. The second element, interest in the inquiry, was based on evidence that the prospective teacher went beyond expectations of the project for the purpose of the class. This connection was measured as high, moderate, or low based on the following rubric. Results are summarized in Table 5.4.

Table 5.4: Rubric setting three levels of engagement in the project. Number and percentage of teachers at each level of engagement with their projects as well as a description of the level is given.

Level of Engagement	Count (Percentage)	Description
High	4 (24%)	their stated connection to the topic affected themselves personally, for example through study of their own race or economic class, or they demonstrated an unusually high interest in the topic that extended far beyond the class
Moderate	6 (35%)	they stated a connection to the topic through family, close friends, or their own (past, present, or future) experiences at work or school, or they demonstrated a very strong interest in the topic through extensive (unassigned) supplemental readings, or repeated requests for assistance outside of class
Low	7 (41%)	they made no stated personal connection nor showed an unusually high interest in the topic

Beliefs about Equity

It is not the purpose of this dissertation to study how the subjects' beliefs changed over the course of their experience, so there was no pre- and post-assessment to document this change. The beliefs recorded here are ones that occurred spontaneously during the presentations at the end of the course or else from the section "Links to Equity" that they were required to include in their final paper. Six teachers also participated in additional interviews and some excerpts are included here where they provide additional insight into the teachers' understanding and beliefs about equity. In addition, teachers were assigned three short essays in January, March, and April which required them to reflect on a particular issue of equity in accountability that was raised in class or a reading assignment. Excerpts from these essays will be used to provide additional illustration of beliefs the prospective teachers articulated about equity.

Project Descriptions

The next sections will describe the inquiry projects in some detail, organized by topic of study (Figure 5.12) and highlighting the pre-service teachers' beliefs expressed about equity, motivation for their study, and statistical evidence presented. At the end of each project description, a table summarizes four elements to enable the reader to see potential interactions at the individual level: the teachers' posttest results (quartile performance relative to the class, from Chapter 4), Fathom behavior (from Section 5.2), statistical use in the final inquiry papers (Table 5.3), and level of engagement (Table 5.4). In addition, each teachers' ethnicity (W=White, H=Hispanic, B=Black) and gender (F=female, M=male) will be given. Overall relationships between level of engagement and the other three variables will be examined quantitatively at the end of the chapter (Section 5.4).

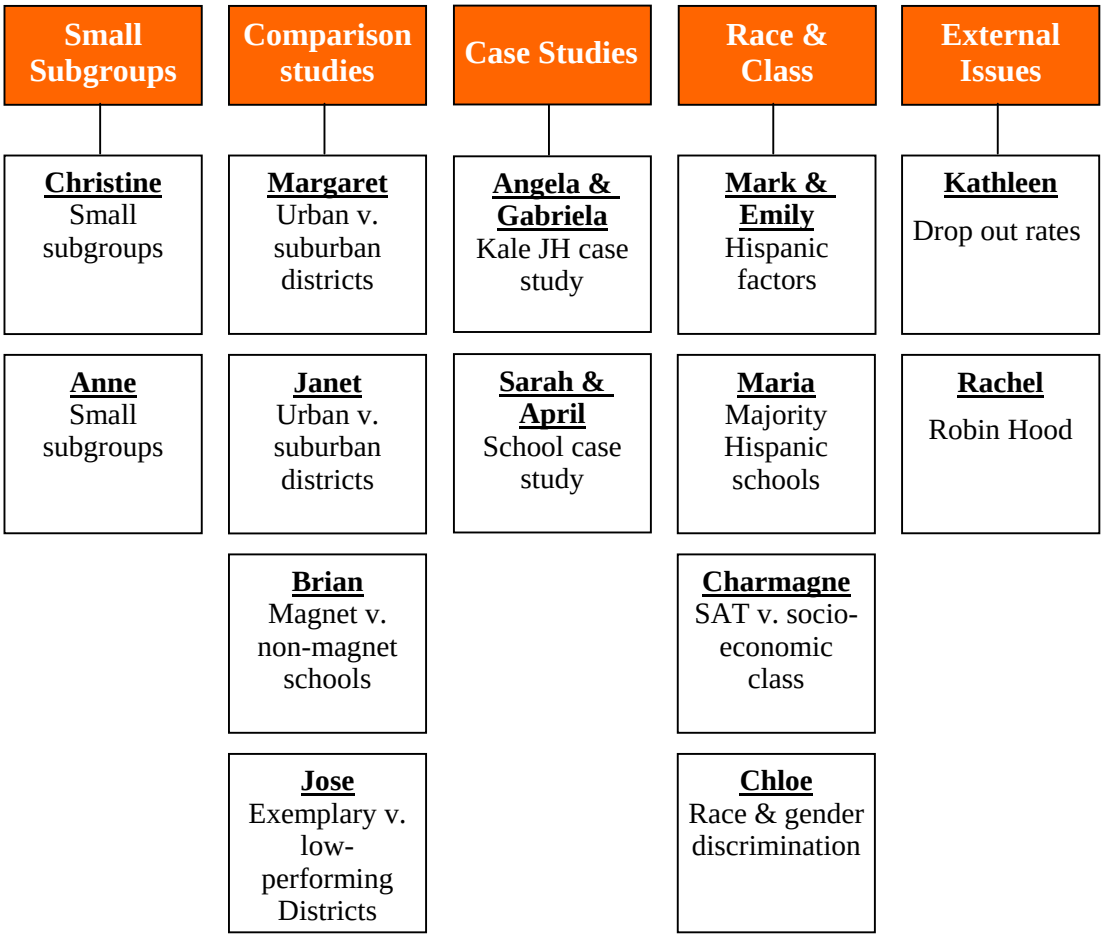


Figure 5.12: Subjects listed by topic of inquiry project

5.3.2 Concerns about Small Subgroups

Although most of the prospective teachers chose a topic that was connected to their personal experience, several chose topics that reflected personal concerns that they had developed about particular issues of equity. They expressed a variety of concerns about issues of equity during the inquiry process that came from analysis of data of school case studies, class discussions, and readings assigned during the course. Below, I present projects which were motivated by concerns expressed by the prospective teachers.

In the first two cases, that of Christine and Anne, both prospective teachers examine the additional risk of being rated low-performing that schools with diverse populations are exposed to. Their inquiries come out of case studies that were examined

during the course of schools, each with a relatively small minority subgroup, that were rated low-performing when this subgroup fell below 50% passing. If one considers a subgroup of students as a sample from a larger population (either of students in the school or of students in this subgroup who pass through the school over the years), one interpretation of this risk is that passing rates based on a small sample are likely to exhibit greater variability than a larger sample would from the same population. Therefore, if a population had a "true" passing rate of 55%, a small (random or representative) sample of the population would have a higher probability of falling below 50% passing than a large sample (Confrey & Makar, in press). This idea of variability of small samples is related to the Central Limit Theorem; the same concept was assessed in the Hospital question on the pre-post test and described in Chapter 4. It cannot be assumed, however, that because Christine and Anne chose to examine this topic that they considered the statistical concept of variation in their choice. It is possible that because we examined more than one low-performing school with a small subgroup, that they considered the low-performance of schools with small subgroups simply a commonly occurring phenomenon, without considering variation.

Christine

Christine expressed concern about the reactive nature of the solutions proposed by schools discussed in the course that were labeled low-performing. She wrote about one troubling case of a low-performing local school (Kurtz, 1999) in her final paper:

> Rather than single out the individual students who required extra help to pass that TAAS test, the school chose to single out the entire ethnic group with the hope that the African-American students who did not need additional help to pass the test would serve as role models for their fellow students who did need help. Although the program was eventually expanded to include a few Hispanic students, the basis for the tutorial groups was race, an unconstitutional and racial practice of tagging students. … By narrowing the participation of students to only one ethnic group, McCallum has sent a message to other schools with a similar problem that the African-American students are to blame for their poor rating, a kind of deficit thinking. … While McCallum's tutoring session may have been responsible for the improvement of the following year's scores, it is not the solution to the true problem at hand: Why were African-American students the only low-performing group at McCallum? (p. 4-5, Final Paper).

She contrasts this reactive response with a more proactive one taken in another school case study, where the focus was on raising the teacher effectiveness rather than

punishing students: “The Glendale community saw the inevitable low-performance red flag on the horizon, but chose to prepare all of its students for the challenge by preparing its instructors” (p. 4, Final Paper).

Christine’s final project reflected a concern she expressed about the treatment of small subgroups and also the risk that having small subgroups posed for schools. She noticed that in many of the case studies of schools with small ethnic subgroups, the schools undertook questionable strategies to raise their performance rating, for example by focusing on raising scores of students near the passing level, mandating tutoring for an entire ethnic group (Kurtz, 1999), or reassigning students to special education (Confrey, 2003; Confrey & Makar, in press). On the other hand, she saw the potential for subgroups to be neglected that were too small (under 30 or less than 10% of the population) to count towards the schools’ accountability ratings. “Why bother helping a subgroup prepare for the TAAS if you do not expect them to count” (p. 3, Final Paper). In either case, vigorous focus or neglect, she felt the treatment towards these groups was unfair and wondered whether these schools could have predicted their low-performance rating based on the size of their subgroups. Since we had discussed several schools during the semester with small subgroups, she also wondered how common it was for schools to have subgroups that were close to 30 in size or close to 10% of the population tested and conjectured that schools with slightly fewer than 30 in a small subgroup (or less than 10%) would have been labeled LP had they had a few more students in their subgroup. In addition, she predicted that schools with subgroups slightly over 30 (or slightly over 10%) would also be likely to be labeled low-performing (LP). Her argument was that these schools would not expect their small subgroups to count and so the schools would not prepare them for the test.

For her sample, Christine chose ten public (non-magnet) high schools in the local school district and examined the passing rates of minority subgroups at each of these schools on the math portion of the TAAS exit exam over four years: 1998 – 2001. She clarified that after the year 2000, the State required all subgroups to report at least 50% passing in order to avoid low-performance; before 2000, the expectation was lower, requiring 40% passing for all subgroups. She presented the results of her inquiry in a very large table, displaying for each high school and each year the rating they received and for each subgroup the count, number passing, subgroup percentage (of all students tested), and percent passing of each subgroup. She then highlighted those

schools that fit her criteria of having fewer than 35 students in a subgroup or having a small subgroup that made up 5-15% of the students tested. Of the ten high schools she examined, three fit her criteria of having small subgroups and were either labeled low-performing or would have been so labeled with only a few more students in these small subgroups. In addition, she noted, this number would jump to six high schools if she had used for all years, the more conservative 2000 performance criteria of 50% passing requirement.

Christine's choice of a sample was quite small and not at all representative of the state. Although a larger sample, more representative sample was available to her, she decided to focus on only ten schools. The choice of data, from a local district, was also familiar for Christine. This may align with other findings of research on children's use of data that shows they frequently focus on attributes from data that help the students retain the original identities or referents, thereby keeping the data more personally relevant (e.g. Lehrer & Schauble, 2000a; 2000b). As will be shown below, several of the other teachers chose "familiar" or small (under ten) samples leading me to think that as an instructor I did not emphasize the inferential power of using a large, random (or representative) sample, particularly given the ease of handling data with Fathom.

Christine's project did not take advantage of the level of statistical knowledge she carried into the course, displaying her data only in tabular form and with no analysis beyond reporting counts and percentages even though she had previously taken a statistics course and demonstrated better understanding of variation and distribution on her posttest (top quartile) and in interviews. Because she did not include any mention of variation in her analysis, it is assumed that her concern of increased risk of low-performance for schools with small minority subgroups was due to the relatively high occurrence of this phenomenon in the cases that we examined during the course, rather than her attention to the concept of variability of small samples.

	Posttest		Fathom Behavior		Statistical Use Level		Engagement	
Christine (W, F)	√	Q4		Wonderer		5		High
		Q3		Wanderer		4	√	Moderate
		Q2	√	Answerer		3		Low
		Q1			√	2		

Anne

Similar to Christine, Anne also examined low-performance issues related to sample size. Also like Christine, Anne was one of the stronger students in the class in terms of her statistical performance. She had previously studied statistics and earned the top score on the posttest. Her project was the strongest in the class in terms of statistical evidence and the only one that made no use of empirical evidence (except to estimate parameters for some of her simulations).

Anne's approach and engagement in her inquiry topic was quite different than Christine's, however. Like Christine, she had been also been intrigued during the course by the courses of action and remediation strategies that schools took to combat low-performance ratings. She wondered how these actions might play out theoretically in simulations. For example, she wondered whether schools with students with diverse performances were at a higher risk of being labeled low-performing than those who displayed more homogeneous performance, even if the mean performance of these two groups was the same. She also wondered whether she could create a statistical model in Fathom to measure whether the strategies used by schools to combat low-performance actually made much difference in their likelihood of becoming low-performing. The approach for her analysis was influenced by the analysis conducted in a reading assigned in class of a school case study (Confrey & Makar, in press) and our subsequent discussions, supported by statistical analyses of the data from the case and related simulations (also described in the study), of an urban high school's reaction to becoming low-performing. A longer description of this case study is given in the introductory chapter (Chapter 1) and in the paper cited above.

To set up her analysis, Anne first created a measure that she termed a school's *low-performance* (*LP*) *risk*. LP risk was determined by drawing a sample from a normal distribution with a mean of 70 (the passing standard on TAAS) and standard deviation of 8 (estimated using empirical data) and calculating the proportion of students that fell below the passing standard of 70. She then created a simulation that drew 100 random samples from a normal distribution ($\mu = 70$, $\sigma = 8$, n varies), calculated this proportion, and found the percentage of samples that were below 50% passing. For a homogeneous (that is, $\sigma = 8$), normally distributed sub-population with $\mu = 70$ and 32 students, she calculated the LP risk to be about 30%.

Anne found that as the standard deviation of student performance increased (keeping the theoretical mean constant), the LP risk also increased, and displayed what appeared to be a logarithmic relationship between diversity of student scores and the school's LP risk, with the greatest change occurring when the standard deviation was less than 10. She also found that for normally distributed, homogeneous populations with a mean of 75 (instead of 70), a school's LP risk dropped to almost zero, while when the mean was 65, the LP risk increased to nearly 100%, "so that kind of puts the parameters on what we are testing. It's about 10 points" (13:10, Final Presentation).

Anne also used simulations to test three possible strategies that schools might take to improve their students' performance: (1) Refocusing resources on the lowest performing students; (2) treating the "bubble kids" (those students who are nearly passing); and (3) creating (or not creating) a program to assist students most at risk of dropping out of school. She found that the mechanism of the accountability system, under a given set of assumptions, was encouraging schools to act in ways that would raise their passing rates at the expense of overall student learning. For example, if schools focused their resources on those students who needed it most, those who performed the lowest, they would dramatically increase their LP risk. On the other hand, by allowing the lowest 1% of students to drop out, a school with an expected mean performance of 70 would cut its LP risk nearly in half. Alternatively, by focusing resources on the bubble kids, and thereby decreasing resources slightly for other students, she found that schools could lower their LP risk to under 2%.

Although Anne's inquiry project was the most sophisticated in the class in terms of statistics, it lacked an expressed emotional engagement with her topic with respect to equity. Her final paper, over 60 pages in length, contained only a half page description of how her project was linked to equity (most of the subjects wrote about two pages of a twenty-page paper). Other elements in her conclusion pointed to the difficulty she had finding connections to equity and perhaps revealed a somewhat elitist perspective. For example, her "best case" scenario described a situation in which students would be "offered equal opportunities to make progress toward higher achievement on the test" (p. 15, Final Paper). This showed that her view of equity may be one of equality of inputs (Lynch, 2000), where equity is measured in terms of equality of resources and opportunities given to students without regard for previous resources and opportunities or level of test performance. Her "worst case" scenario described not inequitable

treatment of students, but fragmentation of programs by what she described as “bizarre strategies” schools used to improve their passing rates and lower their LP risk, for example by not assisting students at risk of dropping out or refocusing resources on “bubble kids”. Finally, she indicated strong negative feelings about the emphasis the test places on the “mediocre middle” neglecting the more gifted students, a group she frequently expressed concern about during the course.

Although Anne did not articulate a strong personal connection to her topic, she displayed a great deal of enthusiasm and engagement in her inquiry of equity in testing through the mathematical models she created. Her inquiry displayed a comprehensive and systematic investigation, through her simulations, of the strategies schools used and the unintended consequences of the accountability system. It should be noted that Anne was a post-graduate student with a degree in mathematics from a nationally ranked mathematics department. Her paper was very clearly laid out with strong links between her goals and the evidence she used to meet them. Her simulations were extremely well documented with appropriate appendices to improve the readability of the paper. She came in several times during the process to have me look at her simulations and ask for guidance and clearly spent many more hours on her paper than was expected. She even submitted a copy of her paper to a journal to be considered for publication. Her inquiry provides an example where her engagement with the statistics in her inquiry led her into a deep analysis of issues of equity.

	Posttest		Fathom Behavior		Statistical Use Level		Engagement	
Anne (W, F)	√	Q4	√	Wonderer	√	5	√	High
		Q3		Wanderer		4		Moderate
		Q2		Answerer		3		Low
		Q1				2		

5.3.3 Comparison Studies

Comparison studies were the most common type of study conducted by the prospective teachers in the course, with 40% of the projects involving group comparisons. Groups compared included: low-performing v. exemplary schools, magnet v. non-magnet campuses, Black v. White and male v. female student

performance, minority performance in ethnically homogeneous v. diverse settings, and finally, performance and characteristics of urban v. suburban districts.

Four teachers did their inquiry projects in this category. Margaret and Janet each compared urban and suburban districts, Brian compared magnet and non-magnet school characteristics, and José examined commonalities and differences in resources for districts rated Low-performing versus those rated Exemplary by the state. Margaret and Janet both expressed interest in comparing urban and suburban districts because although they were more comfortable with a suburban setting, they wanted to teach in urban schools and indicated an interest in investigating the 'urban myths' they were hearing from friends and family, or the press. Although their topic was similar, the approach they took in their inquiry was not.

Margaret

Margaret, an older student, did her project looking at the difference in performance and characteristics between urban and suburban schools. Because she was interested in teaching in an urban school, she wanted to learn more about the problems of urban decline she had heard about. In choosing her topic, she did a great deal of preliminary reading to look for ideas as well as going through a lot of summary data that she found on the web waiting for something to jump out at her.

> So I went and just looked at kinda all the data at a real high level, … just all the different information. And I was hoping that something would just pop out at me that I would want to study. Well, it didn't. (Focus interview, Apr 24 2003)

Margaret had attended an evening panel discussion on equity and testing made up of several scholars at the University of Texas who studied this area. She articulated during an interview that she felt some of the claims made by one professor about urban schools losing their curriculum from the pressure of teaching to the test, were unfounded. In addition, as she did her preliminary readings for the inquiry, she was bothered by the reports she read from experts in the field: "For everything I found saying one thing, I could find an article that was the exact opposite" (Margaret Presentation, April 30 2003). She also expressed frustration during a focus interview (April 24 2003) that she felt researchers were often biased in their reporting and wanted to see for herself what the data would tell her. "I was going to try and see if urban SAT scores and ACT scores had increased. So that might not disprove [that urban schools

were losing their curriculum], but it would at least prove that they are learning something" (08:06). She also expressed that experts may tend to only use data that supports their case and wanted to see the raw data for herself. When asked whether she trusted the data she was finding for her project, she replied:

> I mean, this is, this is pretty, this data, I trust, and right or wrong. There is not an opinion here. It's just raw data. … It's not trying to make a point. It's just delivery. I tend to trust this type of data. And, you know, I have to not question TEA. … I am the type that would love to see all the raw data behind the data to see how they came to it, but you just have to trust it at some point. (16:11).

Margaret summarizes her inquiry topic in a PowerPoint slide during her presentation:

> So I wanted to see, is it really true that suburban always, or almost always outperforms urban? …Is the gap between the scores getting better or worse? … and is there any obvious cause for this gap? (PowerPoint Slide 2, Final Presentation)

In her presentation, Margaret displayed a longitudinal line graph of TAAS passing rates 1997 – 2001 comparing urban and suburban passing rates [she does not say for which subject or grade she used] along with a table with the same information, as well as the "delta" (difference in passing rates between the two groups) for each year. She notes that the gap between the two is decreasing, "but everything you read about, it's only improving because they are teaching to the test, so I can't say why it's improving, but it does look like it's improving" (01:05, Final Presentation).

The sample she drew to compare urban and suburban districts consisted of all ten urban and 63 suburban districts in Texas from data she downloaded from the TEA website. She doesn't explicitly state a conjecture, but takes a very exploratory stance in her inquiry, expressing that she is interested in comparing everything that is similar and everything that is different between the two district types. Her evidence consists of a set of very large tables comparing percentages for each in three major areas: teacher data, financial data, and student data. From the 50 categories she lists in her table, she highlights four particular areas of similarity (with no criteria stated to determine "similar") between suburban and urban districts: (1) number of students per teacher—15.3 vs. 16.1, (2) percent of staff that are teachers—52% vs. 50%, (3) total expenditures per student—$3402 vs. $3507, and (4) attendance rate—96% vs. 94.9%, respectively.

The most notable differences between urban and suburban districts, she notes, are in student demographics. To explore these demographic differences further, she

compares the urban and suburban passing rates disaggregated by ethnicity and economic status by displaying longitudinal line graphs with accompanying summary data and shows that the improvement for African-American, Hispanic, and Economically Disadvantaged students is greater in urban districts than in suburban. Her study, she says, suggests that the reasons for differences between urban and suburban passing rates are NOT because of large differences in teacher quality, finances, or attendance, as she had often heard. She admits that teacher quality was measured by salary, experience, and education degrees obtained which may not necessarily measure quality, but it was all she could find that was measurable. She speculates that the differences in performance between urban and suburban schools might be due to discrimination, parental support, and testing bias.

Margaret, in her comparison of urban and suburban districts makes the assumption in her final paper that if passing rates in different community types are systematically different, there must be differences in the education children are receiving.

> I wanted to further investigate what might be causing this *perceived* inequity in the quality of education being provided. … I've made the assumption here that because there is a gap in passing rates, the education received is somehow different (p. 3, italics in original)

Margaret's assumption that the passing rates at the school could be used to monitor inequities experienced by students and educational quality may point to a belief in equity as equality of outputs, as described by Lynch (2000).

Like Christine, Margaret also did not take advantage of her statistical knowledge to examine issues of distribution or variation in her study, even though she had the second highest score on the posttest and had data from all 73 urban and suburban districts. She was both capable and had the data to compare the two school types in various categories by examining distributions. Rather, she chose to rely on tables and simple line graphs to make comparisons based on summary statistics. Margaret, an older student, had also previously studied statistics and considers herself a "data person"—that in her employment she had extensive experience working with data, making forecasts and budgets.

	Posttest		Fathom Behavior		Statistical Use Level		Engagement	
Margaret (W, F)	√	Q4		Wonderer		5		High
		Q3		Wanderer		4	√	Moderate
		Q2	√	Answerer	√	3		Low
		Q1				2		

Janet

Like Christine, Anne, and Margaret, Janet was a strong student, scoring in the top quartile on the posttest, and had previously studied statistics. Like Anne, she was also a post-baccalaureate student, having completed her bachelor's degree at another university. Janet had been bothered that most people seemed to conclude that the reason urban schools generally performed worse than suburban had to do with economic factors, and felt that her family had discouraged to her to teach in an urban school. She was able to make her project personal by investigating a question that she said had troubled her for some time: "But is that the only reason that suburban schools … are able to earn more frequent Recognized and Exemplary statuses" (p. 1, Final Paper)? That is, was there really a difference between the performance of urban and suburban schools? If so, could this difference be explained by economic factors or were there other issues that could be identified in the data? She wanted to find a way to justify her choice to teach in an urban setting and sought a way for the data to put to rest the story of urban plight described by family and the media.

Although Janet conducted her inquiry using the same topic as Margaret, comparing districts from different community types, she approached her comparison much differently. Using the same data set of 63 suburban and 10 urban districts that Margaret used, she first looked at dot plots of the distribution of passing rates for these seventy-three districts, split by community type. She found the median difference in passing rates between these two groups to be over 10%. She used the scrambling feature of Fathom, which simulates a permutation test, to determine how likely a 10% or greater difference would be under the null hypothesis that there was no real difference in passing rates between urban and suburban districts. By scrambling the collection 100 times, she found that in no case was the difference in median passing rates as great as 10%, concluding the difference to be "statistically meaningful".

She then compared the percent of students in each district that was categorized Economically Disadvantaged (EcD) by displaying this data as a dot plot, split by urban and suburban community type. “If it is proven that urban districts still perform differently from suburban districts within these blocks, one can look to other characteristics of the community type to which differing performance on TAAS might be attributable” (p. 1, Final Paper).

She found that although there was a lot of variation in suburban districts, *all* of the urban districts had over 35% of their students eligible for free or reduced lunch. She decided to compare the performance of just those districts that had over 35% of students labeled EcD, what she termed “blocking”. This meant keeping all ten urban districts, but removing about 30 suburban districts that had fewer than 35% of students EcD in order to put them “on a level playing field … Because I just wanted to see if economic status is equal between them, are they still, is there still a disparity in their performance” (0:51:30, Final Presentation). Displaying the result of the scramble as a dot plot, she shows that:

> Only 2% of the time could natural variation explain a difference in medians as extreme as 6.2% [the difference in median passing rates for her “blocked” subset], if we assume a null hypothesis that the schools are performing fundamentally on the same level. So, the difference in medians may be assumed to be meaningful in this case (p. 4, Final Paper).

Janet admitted during her presentation that she wanted to “disprove” that urban districts lagged behind suburban districts purely due to economic factors by showing that if they had similar numbers of students from poor households, that their performance would be similar. Because there were no suburban districts in the state with high levels of students designated Economically Disadvantaged, she could only partially test her hypothesis, comparing districts with similar, but relatively low, percentages of poor students. This would allow her, she stated, to seek other explanations for the gap, like teacher turnover rate. In addition, instead of focusing on the economic factor—one that was out of a teacher’s control—she pondered whether the difference in performance meant that urban parents were working longer hours and not able to monitor students doing homework or whether there might be less of a push to go to college, areas where teachers could in fact provide assistance and encouragement.

Narrowing the range further to districts having between 70% and 80% of their students labeled EcD, Janet displayed a scatter plot of EcD and percent passing of the seven districts in this range, ignoring community type, finding the correlation to be 0.71,

> indicating that there is a strong moderate to strong linear relationship between these data points. This lends credibility to the theory that there is inequity in performance between schools with dense populations of economically disadvantaged students and schools with fewer economically disadvantaged students, no matter the community type of the district (p. 5-6, Final Paper).

She goes on to make the same comparison for districts on the other extreme, with 30-40% of their students EcD; this time the correlation coefficient is 0.51, for which she states "the data varies wildly along the line" (p. 7). She does not mention whether on this restricted range the reduced sample size will have an effect on her findings or how restricting the range may affect the correlation coefficient. She then goes back to the whole collection and correlates the passing rate with economic status for all schools, not considering the community type (urban or suburban). She displays the scatter plot and concludes emphatically that there is "very evident association!" She notices that, among schools with higher poverty levels,

> there is a great fanning out, whereas the data points for the disadvantaged districts are more tightly clumped. This shows that for schools with few economically disadvantaged students, test performance is consistently good. The fanning out on the left side of the graph leaves open a window of possibility for the truth of my conjecture. Though plenty of evidence has accumulated in this study to help attribute standardized test performance to the economic background of students, this fanning out proves that there is a possibility for excellent performance despite high numbers of disadvantaged students, so factors other [than] economic background of the students must be contributed to this effect (p. 7).

She then points out three possible scenarios for three distinct high-poverty districts, all elements within district's control: crowded classrooms in the lowest performing, small classes and high parental involvement in the highest-performing poor district, teaching to the test in a middle-performing poor district.

Janet's choice of examining equity, by comparing economic groups, focuses on the concept of equity as equality, as described by Lynch (2000). Note Janet's use of the word "inequitable" from her final paper:

> Many people would immediately assert that comparing TAAS scores between these districts is inequitable because of the wealth of the suburban districts in comparison with the large percentage of urban students of limited economic means (p. 3).

Her investigation, to see whether districts that had equality of inputs would tend to exhibit an equality of outputs, indicates a potential belief that monitoring equity begins with an examination of equality. Once Janet established a set of schools with equal distributions of socioeconomic levels, she went on to look at other "input" issues such as teacher characteristics (turnover rate, salary, years of experience, advanced degrees held). She ends her paper with, "There are a multitude of other community-centric factors that, along with economic disadvantages, could be influencing the equity of standardized testing" (p. 10).

Brian

	Posttest		Fathom Behavior		Statistical Use Level		Engagement	
Janet (W, F)	√	Q4	√	Wonderer		5		High
		Q3		Wanderer	√	4	√	Moderate
		Q2		Answerer		3		Low
		Q1				2		

Out of his own experience as a student in a magnet program, Brian expressed his belief that because magnet students choose to attend schools focused on particular disciplines, magnet schools were havens for equity: "Overshadowed by the focus on these disciplines, racial discrimination and segregation fall by the wayside" (p. 2, Final Paper). And,

> As will be shown by the research, magnet schools are important because they serve to shatter the preconceived notions about equity within the education system. Within many of these schools are havens in which race, ethnicity, and gender cease to play a role in the modeling of student performance. … Within these magnets, students are treated with a fairness and equality between students that is not seen at the average high school level. So much so, that disclaimers regarding equity are made in the mission statements of many such schools (p. 3).

For data, Brian compared ten magnet and ten non-magnet programs across the state on several variables: TAAS and End-of-Course exams, attendance and dropout rates, college entrance tests. In each case, he displays the data in both table form and as dot plots, with means marked. After each graph, he includes a short (1-2 sentences)

summary that stated which mean is higher with no mention of other elements of the distribution. An example of one such description is given below for his comparison of TAAS results in the magnet and non-magnet schools:

> TAAS Scores – In the field of 10th level TAAS scores across all sections, it was found (as shown by the tables below) that students of magnet schools typically held a higher percent passing rate than non-magnet schools, and in fact higher than the statewide average. (p. 5)

Note that Brian did not check to see whether students in the magnet schools entered the program with higher scores. His description of TAAS scores was followed by three pairs of dot plots for each of the three 10th grade TAAS exams (Math, Reading, and Writing), and then a tabular listing of the State averages in each of these disciplines. Summaries for his other variables were very similar. Brian made several claims throughout his paper where (in many cases) he provided either no evidence (e.g., "Magnet school populations often tend to be smaller than those of normal high schools"), or the evidence he cited was an attached Excel spreadsheet containing his raw data rather than any graphs or analysis. This seemed to indicate a belief that the data speak for themselves.

Brian used the Internet to get his sample, ten magnet schools in various parts of Texas, but when I looked up the schools he chose, several on the list were actually alternative schools. He did not choose the schools in any systematic way and the size of his sample also hindered his ability to make any generalizations about his results, although much of Brian's very long conclusion, consisting of several pages of discussion, was based not on the data, but his own experience and beliefs about why magnet programs were better than non-magnet programs. In several instances during the discussion, he used results of individual schools to support his claims. For example, he states:

> What was found to be interesting in the magnet school data was that all ethnicities, on average, performed at or near the same level as one another. … For example, Silva Health Magnet School in El Paso was found to have all ethnicities passing all TAAS exams with a rate above 93% in 2002 (p. 10).

In his paper, Brian states that the largest reason for greater equity in magnets is that students apply to enter the program and attend because they want to be there. He indicates also that the smaller class sizes, admittedly more expensive to fund, are

critical to equity: “When smaller class sizes are filled with goal-oriented students, the result is a school in which equity dominates between gender, race, and ethnicity” (p. 4). This statement, together with his beliefs (stated earlier) that one should not consider issues of race in monitoring student performance, again point to an idealized conception of equity as one that is race-blind, particularly when you consider that his analysis of the data did not examine whether his sample of magnet schools had populations that were similar to those in non-magnet schools. Given his earlier admission that minority students frequently have fewer opportunities to succeed in school than Whites, the absence of this analysis is unfortunate.

Finally, Brian expresses his belief that magnet programs deserve the additional funding that they require:

> By my own personal experience, and by the conclusions of this study, I am happy to have learned that our federal government has also discovered the benefits that magnet schools offer to their students, and are currently employing various financial means to support the creation of these programs[4]. (p. 14-15)

This belief is consistent with one Secada (1994) and Kahle (1996) described from the Post-Sputnik era, where the most equitable thing to do is to allocate resources that would enable the largest return on the minimum investment, or largest “bang for the buck”. This seems to be what Brian is advocating by exemplifying equity with magnet programs.

Brian’s poor choice of sample and focus on individual cases and personal experience indicated that he did not necessarily see the potential of statistical tendency to make generalizations about magnet and non-magnet programs. This is despite the fact that Brian had one of the highest scores on the posttest. His mathematics background was very strong (Brian was a physics major), but his previous experience with statistics was only within his science coursework. Again, this may indicate that as an instructor, I did not put enough emphasis on the potential of a well-chosen sample for making generalizations.

Although he did display distributions of his data, Brian mentions only the means when discussing comparisons he made between magnet and non-magnet programs; he made no use of concepts of variation in his comparisons. It should be noted that Brian was likely not finished with his project. When I interviewed him two days before the

[4] Brian never showed any evidence that he checked if magnet programs receive federal funding.

final presentations began, he had not yet begun work on the paper. In fact, the topic he said he would investigate was completely different (investigating equity issues in data from the Physics Force Concepts Inventory) than the one his paper was about. He did not do a class presentation.

	Posttest		Fathom Behavior		Statistical Use Level		Engagement	
Brian (W, M)	√	Q4		Wonderer		5		High
		Q3	√	Wanderer		4	√	Moderate
		Q2		Answerer		3		Low
		Q1			√	2		

José

A Hispanic male, José was another prospective teacher who did not articulate a personal connection to his study of inquiry. He chose to conduct his investigation on comparing characteristics and funding levels of low-performing and exemplary-rated schools.

> I don't have any PowerPoint or anything like that, but, mine is pretty simple. I just went over low performing schools and exemplary schools and just looked at some trends I noticed. I didn't really have anything I thought of going into it. Really, I just wanted to see if there was kind of difference were out there that I would notice. I took a sample of 5, just a random sample of 5 of 2002 from the exemplary list and the small low performing list. … [I] didn't expect to see anything in particular. But I just wanted to see what I would find (45:11, Final Presentation).

José explained later how he chose his "random" sample: "Yeah, I just randomly just went. I went through the 2002 list and I just clicked on them and brought them out." For his presentation, he stood in front of the class and read summary statistics off of a page of notes comparing averages for each sample of five schools on variables such as expenditures per student, teacher salary and experience, class size, and percentage of students labeled Economically Disadvantaged.

In his paper, José expresses an equity belief similar to Secada's (1994) description of equity as meeting the unique needs of each child, except that he applies it to schools, not children:

> Yet, even with laws and programs that try to create equity among the state's schools, is that enough to help all the schools? Not every school is in the same situation, and each school has different needs to be satisfied. I believe since schools are all in different "boats", they require different amounts of funding to help correct problems at their campuses. (p. 1, Final Paper).

Like Brian and Christine described above, José provides another example of a prospective teacher that did not take advantage of a large, random (or representative) sample for his inquiry, even though the data were available to him. He was asked, during his presentation why, given his access to technology, he did not choose a larger sample than five schools. His response indicated that he had not considered that as being important. His level of statistical use improved considerably in his paper, increasing his sample size to 16 for each school category (chosen randomly in the statistical sense this time) and displaying his data as distributions, with means marked, to compare schools rated Low-performing with those rated Exemplary. In many cases he compared only the mean values, but in others discusses range and spread.

	Posttest		Fathom Behavior		Statistical Use Level		Engagement	
José (H, M)	√	Q4		Wonderer		5		High
		Q3	√	Wanderer		4		Moderate
		Q2		Answerer	√	3	√	Low
		Q1				2		

5.3.4 School Case Studies

Angela and Gabriela

Having volunteered at Kale Junior High School (this is a pseudonym, not the real school name), which houses a math and science magnet school, and seeing the quality of teaching, access to technology, and level of learning experienced by the students in the magnet program, Angela and Gabriela assumed that the school would receive a high rating from the State. They were outraged to find that the school was rated only Acceptable, the State's second lowest (of four) rating levels. Gabriela, in her presentation, expressed how this frustration led them to their topic of study:

> We wanted to find out why, what is going on at the school? Or just basically looking at, it made us think like, that's not fair. We know exactly what is going

> on in the school. So there is something wrong when the school only appears to be acceptable [by test scores] and we know what is going on in it (22:40, Final Presentation).

Angela connected it to her own experience in a high school rated one level above Kale.

> I know from my personal experience that [the school] I came from, it was really nothing compared to the middle school that we work at. And for it, for my school to be Recognized and for this school to be Acceptable is just kind of, you know, it was interesting to see how and why that would even be fair (35:15).

One issue that Angela and Gabriela did not consider is that although their experiences in the magnet program may have exposed them to high-quality teaching and resources at the school, the entire school did not have access to these same resources. For example, students who are not in the magnet program are not allowed to use some of the computer labs at the school. The rating at the school reflects not just the students in the magnet program, but the school overall; the performance of students in the magnet and non-magnet programs are not reported separately.

In their investigation of why Kale was given the rating "Acceptable", they examined the results of the TAAS test at Kale. Results were presented in a table with summary statistics (percentage passing) showing the ratings of Kale along with those of a few other middle schools and junior high schools in the same district, in addition to a chart showing how each campus fared compared to the district. The rest of their paper used tables and charts listing passing rates disaggregated by ethnicity for magnet and non-magnet schools in the district as well as other test data (e.g. passing rates of the Algebra End-of-Course exam), demographic information (e.g. percent of students designated Economically Disadvantaged), and teacher characteristics (e.g. years of experience) about the campuses. They also included a quote from a teacher at Kale with her opinion of why the school was rated only Acceptable.

For Angela and Gabriela, equity was based on an issue closer to validity of the system in rating schools, with concern for how this outcome affected students. For example, Angela and Gabriela focused in their project on their belief in the inequity in the rating system more than once:

> It's not fair for the kids, you know, the test that a kid takes, either on a good day or a bad day, should determine whether or not a school should receive funding. But excellent teachers in it, but that one score is going to tell you whether you get money or not (Gabriela, 23:35, Final Presentation).

And, “Everything at the school is being based on whether the school is good or bad is based on this one test. So we just looked at what we know of the school and it doesn’t seem fair” (30:00). Her partner, Angela, agrees and questions, “Because the test[s] aren’t so accurate, is it fair to evaluate them in that way?” (25:53). Their statements here indicate an extreme, and somewhat incorrect, view of the accountability system. Schools such as Kale, that are rated Acceptable, do not lose funding. And although schools that are rated Low-Performing three years in a row can in fact lose their funding, this does not occur based on a single test administration. After being rated low-performing, schools in fact temporarily receive additional funds to support the extra resources they need to bring up their scores.

Angela and Gabriela also discuss inequity in terms of the how well the rating system responds to the uniqueness of individual campuses:

> A major inequity exists in the evaluation of schools. Additionally, there is the issue that no two schools are the same. All schools have a different population, different teachers, and a different economic status. Since no two schools are equal, then perhaps they should not be evaluated the same” (p. 3-4, Final Paper).

This belief appears similar to Secada (1994) and Kahle’s (1996) description of equity as meeting the individual’s unique needs, even though here Angela and Gabriela speak of the uniqueness of schools, not children. Their descriptions also indicate a weakness in their understanding of one main purpose of accountability, as a system of monitoring schools to ensure that all students are receiving at least a minimal level of education, not necessarily to see if all schools are the same in other ways.

Much of the evidence Gabriela and Angela used for their study was rooted in their personal experiences in the magnet program rather than the data that they found. “From our own evaluation of [Kale] Junior High, we found that our most convincing data did not come from any website or graph, but from our very own personal experiences. After working with the [Kale] faculty, we know that [Kale] is a quality educational facility, and the evaluation of its system is unfair” (p. 6).

Their approach in terms of statistics was not to look at trends or distributions but to focus on individual points. For example, they had intended to go and look at each middle school in the district, individually, and see what ratings other schools had earned, but did not cite a purpose for this. When they did make comparisons to other campuses, it was on a case-by-case basis, although they did compare Kale (and their

other chosen campuses) to the district's overall passing rate. They also made several claims during their paper that they did not back up (even though the data were available), for example, stating that the data can show that schools who are rated Recognized have more experienced teachers. When they did use data, they presented it in tables and charts, using only a small and familiar sample of data, rather than displaying longitudinal trends or comparing Kale within a large sample of schools. Gabriela's posttest was about average for the class (slightly above the median) and she had some background in statistics from her science coursework; Angela, however, had not encountered statistics in previous coursework and posted the lowest score in the class on the posttest, so perhaps she lacked the necessary skill set or did not see the relevance of using distributions in relation to her investigation.

	Posttest		Fathom Behavior		Statistical Use Level		Engagement	
Angela (H, F)		Q4		Wonderer		5		High
		Q3		Wanderer		4	√	Moderate
		Q2	√	Answerer	√	3		Low
	√	Q1				2		
Gabriela (H, F)		Q4		Wonderer		5		High
	√	Q3	√	Wanderer		4	√	Moderate
		Q2		Answerer	√	3		Low
		Q1				2		

Sarah & April

Like Angela and Gabriela above, Sarah and April examined a single school as a case study for their inquiry project, which will only be briefly summarized here. They reported that they chose the school it "caught their eye" when they noticed it had improved its accountability ratings for three consecutive years and wanted to investigate how this had happened. They speculated that the increase was due to improved curriculum, but were unable to meet with school officials to confirm this. Sarah and April did not have connections to the school they studied directly, and it was unclear what the personal connection to their inquiry was except that it was a local school where friends had attended and that closely resembled the one that they had gone to themselves.

They presented the demographic breakdown of the school as a circle graph and then proceeded with a series of tables displaying characteristics of the school over a three-year period such as accountability rating, attendance rate, and variables disaggregated by ethnicity: percentage passing for each TAAS test and drop out rate. Their remaining presentation consisted of tables of summary statistics with comments about why the school received its rating every year. For example, in 2001 the school had missed being rated Exemplary (the highest rating) because of dropout rate of the Hispanic population had exceeded 1%. They felt that because the Hispanic population was small relative to the Whites, the school had been treated "unfairly" rated by the accountability system: "We were kinda like, that's not fair because Hispanics had 10 dropouts and that made up 1.3% and the Whites had 10 dropouts and it only made up 0.5%" (Final Presentation). Their final paper included slightly better use of statistics, with bar charts replacing some of their data tables.

Sarah and April entered their inquiry project with the expectation of using data to uncover why their case study school's ratings rose three years in a row. They had assumed, based on the emphasis in the media, that the ratings rose because of increases in the school's TAAS scores and speculated it was due to a change in curriculum. They found that there were several factors that lead to an increase in ratings, including sizes of subgroups, test performance, and drop out and attendance rates. They also note that the system is more complex than to be able to be measured by a single test. The issue of curricular change that they had expected to uncover was not even available in the data. Other teachers remarked during and after the inquiry projects that the data didn't contain the information they had expected and that there was much more complexity in the system that was not evidenced by data that could be collected.

	Posttest		Fathom Behavior		Statistical Use Level		Engagement	
Sarah (W, F)		Q4		Wonderer		5		High
		Q3		Wanderer		4		Moderate
		Q2	√	Answerer	√	3	√	Low
	√	Q1				2		
April (W, F)		Q4	√	Wonderer		5		High
		Q3		Wanderer		4		Moderate
		Q2		Answerer	√	3	√	Low
	√	Q1				2		

5.3.5 Conflicts & discomfort: Discussions about Race and Class

Several times during the course, teachers wrestled with equity issues of high-stakes testing that were discussed in the course. For example, Christine, in her final paper, expresses conflict that some schools prevent underperforming students from counting in the system by retaining these students in 9th grade or transferring them to special education. "Both of these techniques are effective at raising overall group passing rates, yet unfairly label students for the rest of their lives" (Christine, p. 6, Final Paper). Margaret in a discussion about a case of a school with a small underperforming minority subgroup that exempted one or more struggling students from the TAAS, noted that schools are almost forced to hide students in order to protect the reputation of the larger student body and the school's funding. "I mean, you almost have to think of the greater good. So, ethically, you are almost questioning-, it's weird, it just feels wrong entirely. Cause I could see why someone would do that" (Margaret, Class Discussion) The prospective teachers could see conflict in the pressures schools were under and how the strategies some schools undertook to avoid low-performing status were at the same time unethical and effective. It took several weeks before the teachers were able to acknowledge, as Margaret did, that they could understand why schools would do this.

No other topic of discussion about equity created more tension than that of discussing issues of race. Three reactions prevailed throughout the course, those who: (1) felt issues of race were important to consider when discussing equity; (2) explicitly expressed that issues of race should not be discussed; and (3) evaded the issue altogether or redirected discussion to a factor not involving race (for example, socio-economic status or teacher turnover).

One group of prospective teachers, many of them African-American or Hispanic, openly discussed issues of race. José, a Hispanic male, expressed his opinion about issues of race and socioeconomic class in the opening paragraph of a reflection paper:

> In the Scheurich and Skrla article, Continuing the Conversation on Equity and Accountability, there is a section where they talk about how educators have failed children of color. One section in particular I would agree with strongly. 'Children of color and children from low-income families are overwhelmingly tracked to courses at the lowest level. Once assigned to these courses, students rarely get out. Surely we can see that this is a prescription for failing to achieve

> equity in schooling' (Scheurich and Skrla, 232). This tracking that causes this is based on race and income and occurs in racially diverse schools according to Scheurich and Skrla. I think when this is done it gives the students a sense that they aren't better than other students and I don't believe they have the same opportunities as others placed in the gifted, talented, college, and advanced course tracks" (p. 1, Reflection Paper 2).

In this statement, José agrees with the problem of minority students being more frequently tracked into lower level courses, giving them a lower sense of self-worth relative to success in school and fewer opportunities than students tracked into a college-bound trajectory. In another paper, he recognizes problems faced by minorities in schools:

> Deficit thinking basically involves the more powerful group (school) blaming the victims (African Americans) instead of looking at their own selves. It's always easy to blame others when something goes wrong and in the case of McCallum High School, it looks like they just thought to blame the student subgroup rather than taking a look at their own flaws (Reflection Paper 3)

These statements demonstrate a shift from the perspective than he expressed two months earlier, in the first week of the course, when he denies that problems exist:

> One of the students [in the tape] mentioned that they noticed that it seems to be the minority students who seem to have the most trouble with the TAAS. I'm not too sure as to how that holds, but I'm sure if you looked at some charts and graphs and stats, there would be a way to prove/disprove that statement (p. 1, Reflection Paper 1).

Besides demonstrating here his disagreement with the student who expressed minority students in this school had more trouble on TAAS, José also indicates here an early belief that one could use data to disprove the student's statement, a belief in the ability to "prove" statements with statistics that was echoed by a few other teachers in the course.

The two Black women in the course, however, were keenly aware of the problems faced by minority students in schools right from the beginning of the course. As Chloe, an African-American junior, wrote the first week of class, "minority groups are also held back by this test more than the majority race" (p. 1, Reflection Paper 1) Charmagne, the other African-American woman in the class identified herself "as one of the 'other people's children'", referring to the title of African-American advocate and educator Lisa Delpit's (1996) book *Other People's Children*.

In response to McCallum's program of mandatory lunchtime tutoring of all of its African-American students regardless of previous test performance (Kurtz, 1999), Gabriela, a Hispanic junior, felt that "the program seems to kind of blame the African-American students for the low scores" (p. 1, Reflection Paper 3). Emily, a White senior, felt differently: "The positive results from this experiment [mandatory tutoring of black students] should counter anyone's claim that this is a racist way to solve this problem" (Reflection Paper 3). Gabriela's interpretation of the case reflected a belief that such a program may reveal the school administration's deficit thinking (Valencia, 1997) about its African-American students, whereas Emily's interpretation of the case seemed to be that the end justified the means, advocating that schools take whatever steps are needed to do "what works" rather than consider whether the programs treated students equitably.

Like José, Gabriela had not been confronted with racial issues before and never considered that teacher expectations might be different for minority students than Whites, but could see how it could create a vicious cycle of low performance:

> The most surprising thing that I read was how the teachers and administration of schools in a low socioeconomic level tended to reduce their expectation of the students. I realized that there are many stereotypes of people of color, but I never thought that teachers would see their students as anything more or less than kids trying to learn. The articles discussed how this tendency of teachers is a factor of African American and Hispanic children doing poorly on exams. It all seemed like some form of cycle that all results in educators thinking poorly of their students. The Hispanic and African American students in the class do poorly on the test, the teachers expect less from them and don't push them as hard, and then the students do badly on the test again (p. 1, Reflection Paper 2).

Many of the prospective teachers, all White, argued during the course that disaggregating data and bringing light to racial issues only made problems worse. These views of equity articulated a belief that the world should be colorblind, where all people are treated the same, regardless of race. By categorizing a person by race, and potentially treating them differently due to their race, they argued that this appeared to exhibit a form of racism. For example, Janet, a post-baccalaureate student, wrote in one of her reflection papers, "The state's accountability system demands that we look at ethnic groups within the population, but to address the problems with their performance, based on their ethnicity, seems racist" (p. 1, Reflection Paper 3). Rachel, a senior, also expressed discomfort about examining issues of race in one of her reflection papers:

"What bothers me is that we are even evaluating students' performances by ethnicity" (p. 1, Reflection Paper 3). Brian, a junior, had a similar view:

> One of the biggest concerns of mine that I have yet to understand is: Why are the grades broken down by ethnicity within the schools? As a country, we have spent decades trying to break down the walls of segregation. Now, it seems that our Texas legislature has resurrected to concept as a means by which to isolate ethnic groups based on their performance or lack thereof" (p. 1, Reflection Paper 1).

His tone was different two months later in another reflection paper where he acknowledges that there are problems for minority teachers in schools that need to be addressed:

> [The authors] openly address the fact that racism remains to be a source of contention, even today: 'Although the nature of racial prejudice has changed … the data clearly indicates that children of color do not get an equal chance to be successful in school'. I feel that I have to agree. As much as I hate to see it happening, it still appears that minority children are being held back, not for their lack of skill, but for the stigma handed down to them by their teachers and peers (p. 1, Reflection Paper 2).

In the very same paper, however, he reasserts that race should not be examined explicitly:

> Surely there are methods of solving the equity problem that DON'T involve segregating the teachers. You, my teachers, are probably sick of hearing my opinion on this subject, but I still feel that segregating tests scores by ethnicity is wrong. We say that we want to end the time of racial prejudice; we're not helping ourselves attain that goal. … It seems to me that the solution lies, not in the race of teachers (p. 1).

On one hand, Brian acknowledges that minority teachers frequently do not have the same opportunities to be successful as White children. However, he also struggles to let go of the idea that the world should be colorblind and that by examining race we are regressing to times of segregation. It is unclear whether he believes that problems experienced by minorities can be uncovered through disaggregating the data, or if he feels that the practice of disaggregating data reinforces deficit thinking about minorities and contributes to their receiving fewer opportunities.

The term “children of color”, used by the authors of one set of assigned readings, caused a very strong reaction by several White teachers, probably because of its similarity to the old use of the derogatory phrase “colored people”.

> I took offense to different types of ethnicities being referred to as ‘children of color’ (April, Reflection Paper 2)

> I have a major ethical issue with the term ‘children of color’. All three of the discussion papers use this term when referring to African-American and Hispanic children, and all three stress the importance of eliminating racism in schools and in the education field altogether. Yet the term ‘children of color’ *is* a racist category, in my opinion (Christine, Reflection Paper 2).

Although the prospective teachers had difficulty expressing their *own* beliefs about equity, they did not have difficulty in articulating and supporting the two opposing sides of equity issues with regard to accountability. On the last day of the course before the prospective teachers began working on their inquiry projects, they took part in a debate on accountability issues that were raised in the book *High Stakes: Testing for Tracking, Promotion, and Graduation* (Heubert & Hauser, 1999). The issues in the debate were regarding the use of high-stakes testing to decide students’ graduation, retention, or tracking, and the use of testing for these purposes with English-language learners and students with learning disabilities. Although they were able to choose the issue (from the list above) they were to debate, each of the prospective teachers was randomly assigned a position (for or against the use of testing for a particular purpose or group) on the day of the debate. Their arguments during the debate were in general impassioned, well supported, logical, and often innovative. One important difference in this situation than other times they were asked to articulate or discuss their ideas about equity in accountability was that in this case, the audience did not know whether the position they argued was their own or the opposing view.

What the result of their debate indicated was that it was likely that the teachers were generally able to understand the equity issues involved in testing and could articulate, support their reasoning, and challenge ideas on both sides of these issues well. However, in general, they did not show, based on evidence in class discussions, that they were yet ready to articulate, or possibly even recognize, their own positions on these topics. This points to the importance of giving prospective teachers time to wrestle, reflect, and discuss these issues so that in time, perhaps, their own beliefs about

equity would be able to surface in a public debate. It is unknown how long this process would take, however for some of the prospective teachers, as evidenced in their inquiry projects below, they were already willing to unveil and debate these beliefs in public.

5.3.6 Projects Dealing with Race or Class

Four of the inquiry projects dealt directly with issues of race and class. Three of these were by minority teachers whereas the fourth was conducted by a pair of White teachers, Emily and Mark.

Emily and Mark

For their project, Emily and Mark chose to investigate factors that influence minority groups, namely Hispanics and African-Americans, to perform differently than Whites on TAAS. Their reasoning was that "if we inadequately educate this group, it could result in a [de]stabilized economy and an uneducated work force" (Emily and Mark, 01:20, Final Presentation). In addition, they argued that Texas needs to pay more attention to helping its minority students because they will soon outnumber the White population. With this statement, they displayed a circle graph of the demographic breakdown by ethnicity in Texas.

In order to compare factors influencing performance on TAAS, Emily and Mark chose five large districts in different regions of Texas. They introduced their investigation by presenting the passing rates of Hispanics from the two extremes in their sample, Brownsville and Austin: "So we saw this and we were like, well why is it that, that 86% of Hispanics in Brownsville passed and only 65% in Austin passed? So that is our big question. Why is there so much variation?" (Emily, Final Presentation). After listing several factors that they thought might explain these cases (e.g. ethnicity of the teachers, economic level of the district, school programs targeting specific groups), Emily adds, "it's possible that Brownsville just teaches towards the TAAS test a lot more than Austin."

They premise their analysis with a study they found on the Internet by a Harvard Economics professor who investigated how the race of peers influences student performance. The study indicated that Hispanic students in schools with at least two-thirds of its students Hispanic outperform Hispanics in schools that are fewer than one-third Hispanic. "This is important", says Emily,

> because Brownsville is 97% Hispanic; that's way more than two-thirds, you know, so according to her study, that is going to have a positive influence on how students to on their tests. Um, Austin is only 49% Hispanic, so that could be one possible explanation as to why Brownsville does so well (0:05:00).

Mark adds that Brownsville has the highest percentage (of their five districts) of Hispanic teachers, "so we were noticing that the teachers might have played a big role in it, since they were the highest again in Brownsville." They further suggest that minority teachers who are the same race as their students might be less likely to lower their expectations of minority students, and less likely to practice deficit thinking (Valencia, 1997).

As evidence, Emily and Mark display a table giving the percentage of Hispanic students in each of their five districts, as well as the percent passing on the math portion of TAAS of each of these groups. They seem satisfied that since their sample is consistent with the findings of the study, that this indicates the racial breakdown of students in the district is a major factor of success. They then go on to explore whether the education level of teachers might be a factor in these districts. To their dismay, however, they found the data did not support what they were expecting to find and again generalize from two extreme cases, "Um, Brownsville has 83% bachelors, um, and they did the best, but we couldn't really find any correlation with this. Because if you see Austin, it doesn't even have any teachers with no degree. And it even has more teachers with doctorates and masters than Brownsville, and they did the worst" (Mark, 0:07:00, Final Presentation). Emily and Mark conclude their presentation by summarizing their findings:

> So in conclusion, there is no one factor that influences student achievement. And, um, racial balances of both the students and the teachers do play some part, although it's not always, it's not an exact science. … So, um, and then of course, like we just said the higher education of teachers is not always too correlating in increases in achievement.

During class discussions, Emily and Mark approached issues of race differently. Emily did not shy away from discussing issues of race, usually displaying a pragmatic stance that the end justifies the means and that issues of race need to be addressed because there are clear differences in performance between minority and majority groups, a view of equity as equality of outputs (Lynch, 2000). Mark, like many other White teachers in the class, almost never entered into discussions of race, steering the

discussion towards a different perspective. For example, at the end of their presentation, one of the instructors in the course asked Mark whether a principal with a large Hispanic population should try and hire a Hispanic teacher over a White even if the Hispanic teacher has a different education level. His response was to turn the discussion from one of race to one of expectations, saying it's not that the teacher is Hispanic, per se, that makes the difference, but that Hispanic teachers might hold their Hispanic students to a higher level of expectation. He goes on,

> whereas a White person, you know, I'm not trying to be derogatory here, but like a White person, there is some evidence that they might, just because it is a different race, not hold their standards as high for [Hispanic students], which is maybe part of that deficit thinking.

He recommends instead that the principal try to "get a sense of their personality" and make their decision based on that. After a long pause, Emily disagrees, saying that race should play a role. She uses herself (again a single case) as an example, saying that she doesn't think she will be a successful teacher in an urban district with a large Black population, because she doesn't have much experience with African-American students. She continues,

> *Emily*: You know, maybe a, a male, an educated black male can serve as a much better role model than me.
>
> *Mark*: I actually noticed that through my soccer camps even. I mean, it's just very prevalent. We have a guy who went to the University of Houston. You know, black guy. He can really get on the level with the black kids.
>
> *Emily*: It may not be fair, but it's-
>
> *Mark*: It works.
>
> *Emily*: It's true.

This was the only time that I heard Mark publicly address issues of race. He was clearly uncomfortable as he spoke, with his face bright red.

Mark had previously taken some statistics and Emily took a research methods course concurrently with the study. It was disappointing that they chose such a small sample, focused on only a few cases to make their argument, and displayed all of their findings as tables of percentages. This indicates that they saw little value in using statistics as evidence. The most compelling evidence for them was in the extreme

cases—if the cases fit their theory, then they said the data correlated and if not then the data did not correlate. Both Emily and Mark entered the course with above-average performance on their pretest and made below average improvement on their posttest. In addition, they both entered with high levels of confidence in their understanding of statistical concepts but both made average to below average growth in their confidence level during the course.

	Posttest		Fathom Behavior		Statistical Use Level		Engagement	
Emily (W, F)		Q4	√	Wonderer		5		High
		Q3		Wanderer		4		Moderate
	√	Q2		Answerer	√	3	√	Low
		Q1				2		
Mark (W, M)		Q4		Wonderer		5		High
		Q3	√	Wanderer		4		Moderate
	√	Q2		Answerer	√	3	√	Low
		Q1				2		

Maria

Maria, a Hispanic female, addressed a topic similar to that of Mark and Emily in her inquiry project: whether Hispanic students performed better in districts that were majority Hispanic, majority White, or districts where the population was racially diverse. Like Mark and Emily, she chose a rather small sample, just 11 districts located in various parts of the state. Maria's interest was more personal, however, not just because she was Hispanic, but also because she grew up along the border region of Texas in a school that was 95% Hispanic but had never experienced problems on TAAS. She expressed surprise during the course that minority students were frequently blamed for schools being low-performing and wondered why the minority students in the schools we looked at were performing so poorly. She wanted to find out what might explain the difference in performance between these schools and her own and used the inquiry project to test her theory that it had to do with the demographic breakdown of the district.

> My original hypothesis was that school districts that are predominately, particularly minority that were predominately Hispanic, will have a higher percent passing than when that same group is of a smaller population. Those are

> the things you've had, you've always heard about how they have, they're like a little small population of Hispanic students or little minority group and that caused the school to go low-performing. And then I had been along the border, and I've only really known about schools that were predominately Hispanic. And I never really heard about them going low performing or being in trouble or anything like that. So I wanted to see if there was a difference between that. (16:00, Final Presentation).

Maria examined several aspects of Hispanic performance on TAAS, different subjects and different grade levels, all with dot plots or box plots displaying the distribution of performance (percent passing) split into three categories based on demographics: largely Hispanic, largely White, or racially diverse. Maria's hypothesis was that Hispanic students would perform best in a school with a largely Hispanic population, and she was clearly unhappy with the results from her sample that showed Hispanic students in majority White schools outperformed Hispanic students in majority Hispanic schools.

> And then, I was kind of surprised because the ones that were mixed diversity were somewhere on the lower end and the ones that were predominately White were on the upper end. … I saw that my original hypothesis was not what I thought it would be. I started looking at other factors that could have caused it. And I started looking at student revenue per student because each districts has a different size and I thought they would be different. And then I looked at the schools and districts, how many there were. And then, I started looking at the number of students to teachers, percent of economically disadvantaged, uh, percentage of students passing the classes, and the diversity within the districts.

Maria's evidence for this part of her investigation involved scatter plots (e.g. revenue per student vs. district size marked by category), dot plots (e.g. percent of students eligible for free and reduced lunch), and box plots (e.g. percentage of Hispanic students at schools in each district category). She concluded that the predominantly White schools in her sample tended to be small, wealthy suburban districts whereas the more diverse schools were all large urban districts with lower revenues per student. Responding to a question about how she chose her sample, she says:

> I picked the major cities that I knew I would get good diversity results. And then I was just picking ones that I kinda knew, well I was having help. People were telling me that this district is predominately White, so I was just picking some [that way].

Like Emily and Mark, the method of choice or size of the sample did not seem to be problematic for Maria at this point, yet clearly in both of these projects, the results of their investigations are questionable because of their sample choice. In her final paper, perhaps due to feedback she received about the size of her sample, Maria acknowledges that a larger sample (not necessarily chosen differently) would have been better:

> This investigation is just basically a generalization based on a few districts from the entire state. Given the time allotted for this investigation, it would be better to have a larger sample population than about 3 from each sub district [category] in order for my [investigation] to have more validity. Likewise, it would be interesting to see if these trends hold true in past years as well as future years instead of focusing on just the year of 2001 (p. 14, Final Paper).

A major difference, however, between these two projects was in the way they used the data from their sample as evidence. Mark and Emily relied on reading values in tables of percentages they presented and focused on single cases to make their argument of whether variables were "correlating" (i.e. consistent with their theory) or not. Maria, however, relied on distributions to make her case, and rather than simply reading the data, she interpreted what she saw and drew generalizations from her data, although usually by focusing on the order of magnitude from results in her three categories. In the write up of her project, she again expresses surprise at the results that counter her conjecture that Hispanic students would perform best in largely Hispanic schools. She is not ready, however, to let go of her personal experience that countered the results of her analysis.

> Though I knew these results were only a sample of a population I was mind boggled by the results I had found and that is when I started focusing on other factors that could have caused the results which lead me to get an idea of what is going on in these districts but this led me to more questions and uncertainties.

Maria's reaction to what she found may point to the existence of several emerging beliefs: (1) because she had only a sample of the population, she understands that her findings may not extend to the whole group; that is, although Maria never articulates that her sample was not well chosen, she seems to be expressing doubt whether characteristics of *her* sample, or perhaps of samples in *general*, are necessarily representative of the whole population. (2) Her expression of being "mind boggled" by her results indicates that they contradicted her expectations. This indicates that she *had*

expectations of what her outcomes would look like. Also, the conflict between what she saw in the data and her own experiences pushed her further into her inquiry to search for an explanation. Her experiences here coincide with Dewey's notion that inquiry is ultimately rooted in the need to resolve doubt and that this doubt acts as an internal motivator. Finally, (3) her admission that the search for solutions led her to "more questions and uncertainties" indicates either her frustration, or recognition, that the data and results of an inquiry do not always provide definitive "answers" to what one is looking to uncover. Whether frustration or recognition, this indicates that Maria's previous experiences with scientific inquiry likely did not provide her with the discomfort she is finding in her inquiry here. What is hopeful in Maria's statement is that she is neither clinging solely to her own experience nor accepting the results of her analysis without further probing; rather, she is looking to resolve their conflict and seeking to integrate her results with her experiences.

The next section of her paper, Maria concentrates on the less uncertain process of searching for factors that might explain her results. Here, she methodically goes through several variables to look for patterns that might emerge; the descriptions of her graphs focus primarily on the order in which the mean of each category falls for each particular variable. In each description, she used a clear chain of reasoning to link the outcome of her analysis to the context of the problem, explaining how it relates both to the variable under investigation and the larger question of Hispanic performance. As she goes through each variable, the analysis leads her to further questions and into another analysis. For example, when she examines district size as a possible factor in performance, she wonders how the size of the district might affect the ability of teachers to provide one-on-one attention for students who need it. This leads her to look at revenue issues. Her approach here mirrors her behavior as a "Wonderer" in the Fathom interviews earlier.

	Posttest		Fathom Behavior		Statistical Use Level		Engagement	
Maria (H, F)		Q4	√	Wonderer		5	√	High
		Q3		Wanderer	√	4		Moderate
	√	Q2		Answerer		3		Low
		Q1				2		

The other two teachers who focused on issues of race and class in their inquiry projects, both African-American women, showed very different levels of sophistication and engagement in their inquiries. These two cases are described in detail below.

Charmagne

Throughout the course, Charmagne, a sophomore, had been fairly quiet. She rarely spoke in class and privately expressed concern about her ability to pass the course because of her lack of statistical knowledge. Her poise and confidence changed, however, as we began discussing equity issues halfway through the course. She came in several times for help with the statistics homework during this time and excitedly expressed to instructors that she found the discussions in class about equity very, very interesting, although for the most part she continued not to speak out in class. In her final project, Charmagne chose to investigate the connection between SAT scores and poverty.

> So the problem is, is, um, does the higher education system promote inequity by using the SAT test as a factor in admissions? Um, my conjecture is, um, high SAT scores will be directly correlated to high economic status, thus low economic groups are subjected to inequity (31:00, Final Presentation).

Charmagne expresses concern that lower income students will not have access to SAT preparation programs such as the expensive software or tutorials marketed by private companies, like Kaplan or Princeton Review, that promise a 100-point increase in combined SAT scores. As evidence, she displays a bar chart and table that list the average SAT score for a range of income levels, reported in $10,000 increments. Using Fathom, she displays a scatter plot of schools' mean expenditure per student and their average SAT scores and comments on the strong positive association between a school's expenditure and its mean SAT scores (although doesn't state a correlation coefficient). Showing a data table listing relative income for Whites, Hispanics, and African-Americans, she emphasizes that issues of economic status are also issues of race. Her stance by now is strong and confident.

As she finishes her presentation, Charmagne tackles several tough questions from the audience, more than in any other presentation. April, a White teacher, is somewhat defensive and expresses that all students have access to preparation materials at the local library. After all, she says, "I didn't take a course or anything. I went to the library and checked out books. You know, I mean, did you check out … [if] in certain

communities those resources are lacking?" (39:30). Charmagne calmly responds that she did not examine data on community resources and acknowledges that it would an interesting idea to look into. She then responds to a tough question from a course instructor: whether she thinks this is an issue of *equity* or *equality*, and given that some groups may systematically come less prepared whether she still thinks this is still an equity issue. Charmagne is very emphatic in her response:

> When you can pay $1,000 to get a private tutor, one on one, that makes a LOT of difference! Like going to the library and checking out a book and doing it yourself. That is a lot, there is a LOT of difference there.

Although the evidence in Charmagne's presentation emphasized equity as equality of inputs (income level, preparation resources) and outputs (SAT scores), her concern about the inequities in college admissions criteria displayed a belief in equity as fairness and consequential validity (Messick, 1995). That is, that the consequences of the inequalities in resources and the interpretation of SAT scores led to inequities in college opportunities for low economic groups, and therefore also for minority students. At the end of her presentation, Charmagne admitted taking a Kaplan-like course herself three years in a row to bring up her own SAT scores in high school, raising them by almost 200 points. It's possible that she believed that those courses, and the finances that paid for them, were the reason she was able to get into the University.

In her paper, Charmagne's analysis went into more depth. She used a sample of 100 schools from data downloaded (with help) from the Texas Education Agency website (www.tea.state.tx.us) to examine a scatter plot showing the percent of African-American students taking the SAT vs. the percent of Economically Disadvantaged students at the school. She also shows the same variables in scatter plots for Hispanic and White students at the schools in her sample. The graphs that she displays show, for example, that for schools with greater than 65% of students classified as Economically Disadvantaged, none of their African-American students took the SAT test whereas in schools with fewer than 25% of students qualifying for free or reduced lunch, 100% of their White students took the SAT. Although she at times was confused whether her data points represented schools or students, she was able to use the scatter plots to find interesting relationships between opportunity to take the SAT test and economic status. Using the same variables in a pair of linked dot plots, she finds that for the 17 schools

(out of 100) that did not take the SAT at all, ten had over 50% of their students from low-income families. She struggles expressing herself, but makes a strong point:

> If we focus on the African-Americans, we find that [the schools where] the students … did attempt the test were below the 60 percents disadvantage status. Half of the [schools which had] African-Americans who took the test had less than thirty percent of their students at a disadvantage. The same pattern follows for the other two subgroups [Hispanic and White] (p. 8, Final Paper).

Charmagne's paper made good use of the readings from the course, including providing SAT examples to illustrate very difficult concepts about measurement validity, attribution of cause, and effectiveness of treatment outlined in chapters from *High Stakes* (NRC, 1999) that we read during the course. All of her data results were clearly linked to her conjecture through strong chains of reasoning between her conjecture, choice of data, evidence presented, conclusions drawn, and equity implications that her results had for poor and minority students in opportunities to attend college.

Charmagne displayed an extremely high level of personal engagement with her inquiry. This was evident in her passion about her topic, interest outside of class in obtaining useful data, and personal connection to the topic through concern about inequities for minority students. A year later, Charmagne described a project she was engaged in at a local magnet school to examine and decrease the gap in performance between the magnet (predominantly White) and non-magnet (predominantly minority) students at the school. This choice points to her deep commitment she has to equity and her resolve to act on her beliefs about equity to decrease inequities for minority students in schools.

Charmagne's high level of statistical evidence used in her project was not a result of a strong background in statistics. She entered the course with no previous statistics and clearly struggled with the review assignments designed to provide students an opportunity to practice and extend their understanding of the statistical concepts that arose during class discussions and investigations. Charmagne earned one of the lowest scores on the pre- and posttests, and posted slightly below-average improvement. However, she posted one of the higher levels of improvement in the class in her confidence (top 25%). The topic she chose was one that she was engaged in and was personally meaningful to her. Although Charmagne's analysis did not go to the top level of statistical evidence (level 5), as Anne and Janet above, she did have one of the

highest levels of statistical evidence in the class in her project. She relied on and spoke of association and variation in scatter plots, used linked dot plots to compare relationships between variables, and clearly linked the evidence she used to her conjecture throughout her paper.

	Posttest		Fathom Behavior		Statistical Use Level		Engagement	
Charmagne (H, F)		Q4		Wonderer		5	√	High
		Q3	√	Wanderer	√	4		Moderate
		Q2		Answerer		3		Low
	√	Q1				2		

Chloe

Another teacher who included discussions of race or class as a focus in her inquiry project was Chloe, an African-American junior. Chloe was also part of the focus group of teachers that participated in additional interviews during and after the inquiry projects were conducted. During the final focus interview, Chloe talked about the topic of her inquiry: "My paper is um, is about high-stakes tests and whether or not they are biased. And um, with certain groups." Chloe chose to compare the performance on TAAS of two pairs of subgroups in her inquiry, Black versus White students and male versus female students, because she felt that females and blacks were the most commonly mentioned in issues about high-stakes tests, and she is a member of both groups. "And, um, I just wanted to see if like there really was biasness, I guess." She clarified her interpretation of bias by highlighting three kinds of differences that emerged as issues for these groups: differences in scores, differences in test content, and differences in opportunities to prepare.

> Like if you could tell if, um, male scores or White scores would be higher than, um, the other students' scores based on content of the test and um, based on things that students were able to use to prepare for the test. … And I talked about opportunities in the paper. About opportunities that one subgroup might have and others might not have.

This was a considerable shift in meaning from what she gave initially for her topic in an earlier focus interview a few days before she presented her findings to the class:

> It's just a fact, well, from what I have heard and from other people, there are discriminations, but I'm looking to find proof of discriminations. I'm, me, myself, I don't believe that there really like noticeable discriminations in it, but there may be some. And I am hoping to see, hoping to be able to find out if there are or what they are.

The language in her initial topic choice focused on things she had "heard" about discrimination and bias, and she wanted to find "proof of discrimination". Her language in the later focus interview concentrated more on differences and opportunity, rather than discrimination. She explained the change in her perspective, from having a general impression about discrimination on the tests to thinking more about disparate impact:

> *Chloe*: At first, I thought it was about the same. But then after I gave my presentation … [Dr. Confrey] brought it to my attention and made me think about it more. Made me realize that bias and discrimination aren't really the same thing. Because like, when I wrote this paper I was thinking about biasness, biased tests, and not, not how, um, I don't know how to say it, not how they would be discriminated. Not how they would discriminate people. And I don't know how I switched gears, I don't know when I switched gears.
>
> *KM*: … So, to you is bias closer to the idea of discrimination or closer to the idea of disparate impact, or does it have a meaning that is different than either of those?
>
> *Chloe*: Um, can it be somewhere in between? Um-
>
> *KM*: It can be somewhere in between
>
> *Chloe*: I think it is in between, on the closer end to discrimination. But not as far up as discrimination.

For Chloe, this distinction—from hearing from others about the tests as discriminating against Blacks to showing disparate impact—was critical for her understanding of equity. Although she was still grappling with what discrimination, bias, and disparate impact meant to her, she was cognizant that these were elements she wanted to think further about. In a focus interview a week later, conducted after she had presented her findings and written her final paper, Chloe had clearly spent more time reflecting on this distinction. Her explanation of what equity meant to her included a greater focus on opportunity, validity, and purpose:

> *Chloe*: Equal opportunity on tests, tests being fair to different groups and having everything, reason to be in there. Is that what? Yeah? Reason –

KM: Reason for what being where?

Chloe: A reason for the test to be given…. To make sure the test is really testing what the student should know or what they are being taught in school. Or if it's just some bizarre test that they are just giving students to say we have a standardized test.

From these excerpts, a progression of ideas about Chloe's conceptions of equity and fairness in testing emerges. She began the project thinking that her investigation would uncover discrimination in the tests. She admits that even though this perspective was not consistent with her own experience, she accepted, without further evidence, that discrimination must exist because she had heard that the tests were discriminatory. Her goal, therefore was to "prove" this discrimination, against Blacks and against women. These goals persisted after over two weeks of investigation of the data, even a few days before she presented her findings. Based on feedback during her fifteen-minute presentation to the class, however, she began to reflect on and reconsider what it was she was investigating. By probing how she differentiated the concepts of discrimination from disparate impact, she began to consider the concept of equity in testing more deeply: what it meant, how it was monitored, what might be done about it.

By the time she had finished writing her paper, Chloe's focus had shifted from a general notion of discrimination, something that she might be a victim of, to issues that she might be able to examine more definitively and forge efforts to improve: differences in test scores, differences in test content, differences in opportunities to prepare. Some of these distinctions were just emerging, however, and while she recognized that what she was seeing went beyond discrimination, she was not yet able to articulate what she was thinking. In her final paper, she wrote only a few lines about equity, lacking confidence that she had anything to say. During our final focus interview, when we went through her paper, she spoke further about her experiences over the course of the semester and how these experiences had an impact on her thinking. Validation of these experiences increased her confidence in her ideas and she was then able to articulate, very well, what she was thinking about equity.

Chloe had always been a successful student throughout school, enrolled in the Gifted and Talented program since elementary school, and was frequently one of the only Black students in her classes. Because she never had that much trouble with the TAAS test herself, she was surprised to see such a large difference in average

performance between Blacks and Whites and wondered why there was a gap. In addition, she was usually one of only a few women in her more advanced mathematics courses in school. She said her male teachers didn't overtly show favoritism, "but you can tell, there is something there. That they showed more in with helping the guys do better than the females" (9:34, Focus interview, May 8 2004). At the University, she had recently taken a number of courses that examined Black issues in history, culture, and literature and found the content new and fascinating. She said she had learned about how minority groups have to work a lot harder than the majority group to be on the same level, in order to succeed and to do better than minorities have in the past. She goes on to talk about the struggles of blacks in the 1950's through 1970's.

> *Chloe*: I realized that we are still fighting the same battle that was being fought then, but not as much as, it's not as noticeable but it's still there.
>
> *KM*: So has that been kind of meaningful to you? Was that something that you didn't realize before studying that in English?
>
> *Chloe*: I actually yeah, I have never really, where I am from, I never really noticed like um, what do you call it? Discrimination or racial hostility or anything. And then when I came and I started taking this English class, it was kind of unique because like at the beginning of the semester when the MLK statue was egged [on the University campus]. And we were talking about that and like two weeks before the teacher had asked us if we still noticed racial tension or not. And I had never really paid attention to it until this semester. And this semester has really brought it out in most of my classes. So it was just like something, I don't know. And it was really, and I think that is one reason why I chose this topic, too, because it's really a way for me to dig deeper into different things that will help me when I become a teacher to make sure that I am not part of that group that makes that gap bigger between different sub-groups. Between different groups of people.

In conducting her investigation, Chloe found the inquiry process frustrating. "I have the data and I don't know what to do with it. I don't know where to go from here. It's like, I know what I want to look at, well I know what my question is, but I don't know exactly what to look at and how to use what I am looking at to answer my question" (6:05, Focus interview, April 28 2003). Chloe said that one of the most frustrating parts of the inquiry was the open-endedness of it and that there was no specific method to follow. She describes how she wishes she could work through research an expert in the field had done and then try and replicate the experts' results. Although Chloe was a good student in school, I suspect that her experiences were very

structured, so that she was taught to follow the procedures taught to her without having an opportunity to either construct meaning or experience complex questions that involved uncertainty and multiple interpretations of evidence.

For her inquiry project, Chloe used three sets of data. One set was a sample of data on 50 schools taken from data downloaded during class from the Texas Education Agency (TEA) website (www.tea.state.tx.us). She said she felt that 50 schools were enough data and that if she used more data it would have been too confusing, "boggle up the data and you wouldn't be able to clearly see". The other two data sets she used were both student-level data taken from a random sample of 10,000 student scores in Texas obtained from TEA by the instructors of the course. Student-level data on ethnicity consisted of data combined from a randomly chosen set of 50 White students and 50 randomly chosen Black students. The gender student-level data set was larger, with nearly 500 students altogether. She examined differences in performance at both the school level and student level.

In her presentation and her final paper, she displayed dot plots of percent passing for each pair of comparison groups (White versus Black and male versus female) from the school-level data, and also dot plots of MTLI (scaled score on the mathematics portion of the TAAS) for each pair of comparison groups from the student-level data. From examining the distributions of each group (no means were marked), she found that although neither the school-level nor student-level data showed any differences in gender, the data comparing race had a different outcome. The school-level data displayed a higher percent passing for Whites than Blacks, but the student-level data showed no difference in average MTLI for these two groups. She did not mention or seem surprised that these two data sets appeared to show contrasting results.

In her presentation, it was clear that she struggled with how to interpret the data and analysis she conducted, although her understanding improved when she turned in her paper a week later. She had attended an optional two-hour workshop that had been made available to the class in using a procedure in the software to compare groups (a permutation test, called *scrambling* in Fathom), so perhaps initially she felt that she was supposed to use this procedure in any group comparison. It was not clear in her presentation that she understood what a permutation test would tell her, although she was clearer in her final paper where she explained how she used the scrambling distribution to "see how graphs would look if there really was no preference to whether

a black student or white student took the test, or if a male or female took the test" (p. 3, Final Paper). She used the original difference in scores on each sampling distribution to, "see how many students were pass[ed] the 'normal' spot and how this changed my prediction from the beginning. I realized that I would not be able to tell if the different groups were being discriminated against just by looking at these graphs" (p. 4).

A positive outcome of her inquiry seemed to be that although she went into her investigation seeking to prove discrimination, she could see that the test data really did not measure what she thought it would. She had initially interpreted the difference in performance as meaning there was discrimination in the test, but later realized that she would not be able to assert discrimination based on differences in test scores. In addition, since she had "heard" there was a difference in performance between Blacks and Whites on TAAS, she seemed to be expecting that this difference would appear as two non-overlapping distributions. This assumption cannot be confirmed as I did not ask her to draw what she anticipated seeing, but frequently in interviews where subjects were asked to compare groups, they indicated surprise that distributions with different means had most of their values overlapping.

In her final paper, Chloe compared not summary statistics, like many of the other teachers, but distributions (Figure 5.13):

> The African American Scores percent passing scores on the Math portion on the TAAS test are spread out just like the student individual scores. This graph shows more of a variety among the students to having most of the people in the group pass to having a small amount pass. The whites on the other hand are all together at the higher end of the percent passing scale which shows that the majority of the white students that took this test passed and hence their school has a higher percent passing average than the black students. Looking at the male and female graph we can see that both the male and female have percent passing rates that are spread out over a wide range. This shows that the percent passing of each group in the school also cannot be pinned to one place. This graph shows that the range of grades in these schools on the test is really wide.

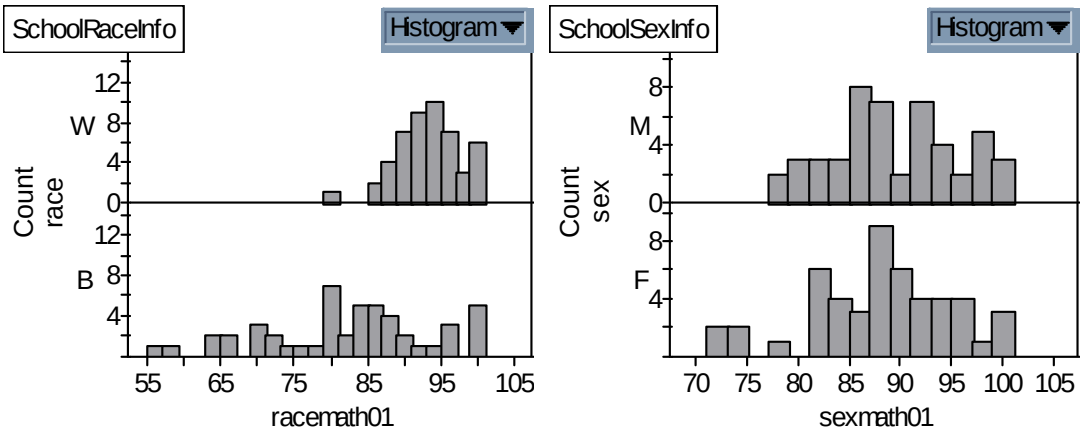


Figure 5.13: Chloe's histograms comparing the distribution of percent passing rates at 50 schools between Black students and White students on the 10th grade mathematics portion of the TAAS (left) and between males and females (right).

Chloe's interpretation of the school-level distributions comparing percentage passing between genders (Figure 5.13, right) indicates that she sees the overlap and similar variation in scores between the genders as meaning that scores "cannot be pinned to one place". Note in her graphs that she did not plot any measure of center, but focused in her descriptions on the distributions of the data. She did use summary statistics, however, to compare the distributions when she scrambled the data by calculating the difference in medians or difference in passing rates between the groups. In interpreting the sampling distributions she generated from her permutation tests, she showed that her understanding of sampling distributions was still developing. Although she understood that the purpose here was to compare the measures she obtained from her original data, she struggled to interpret the meaning of points in her sampling distribution. For example, under the null hypothesis that the genders performed equally, the sampling distribution of median differences should have centered near zero, but all of her values came out negative (due to an error in syntax in creating the Fathom formula). She did not see a problem with this and interpreted it to mean that the girls always outperformed the boys (Figure 5.14).

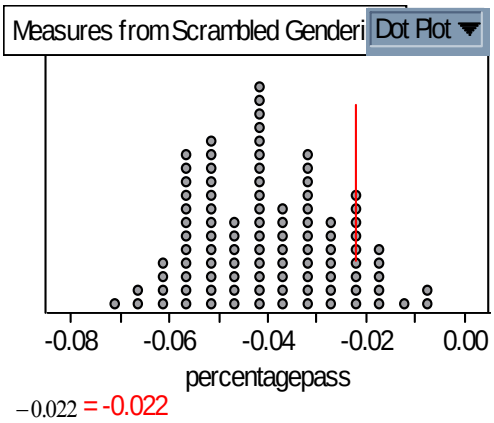


Figure 5.14: Sampling distribution generated from Chloe's scrambling procedure comparing passing rates in student-level gender data. The value of 2.2% calculated as the difference in the original student-level data is marked.

Chloe's conclusion in her paper sums up a number of issues for her.

> [My] findings really did not help with my [intended] investigation because we can not just look at the simple data to see if the tests were biased or not. I could not even tell if there was discrimination to the minority groups that were involved in taking the test. The only thing that I could see was that there were differences in the scores of the groups as a whole and to me that difference is big enough to look into the resources that are given for preparations to the test and how each school district handles their schools whether they are rich, poor, male, female, white, or black.

Chloe saw that the data from test performance really cannot tell her whether or not there is discrimination between the groups, only that the difference means that issues of discrimination and opportunity to learn need further investigation. She felt her findings did not 'help' her investigation because the data did not clearly tell her whether or not there was discrimination. I can't tell from this if she means that the data did not give a single outcome to support or refute her conjecture, or if she is saying that this data *cannot* give her the information she needs to determine discrimination because the data do not measure this. On the other hand, she does sum up what the data *does* tell her: that differences exist, big enough ones that they need to be looked into in terms of resources, preparation, and "how each district handles their schools whether they are rich, poor, male, female, white, or black" (p. 7).

Although Chloe struggled interpreting the results of her analysis, she indicated through her interviews, presentation, and paper that she was very interested in

uncovering tensions she felt between her own experience as an African-American woman and the results she "heard" presented by experts and in class about discrimination and differential performance on TAAS. She was the only teacher who had never studied statistics that chose to use sampling distributions which took a great deal of courage to undertake. She also indicated in her descriptions that she valued comparing not just summary statistics, but also distributions of performance. Chloe had stated in an early interview that she doesn't really understand that summary data might give her different information than distributions of data, and expressed in her final paper that when beginning her inquiry, "I was not really sure how the data would help me out, but it seemed like a good starting place" (p. 2). Initially, she thought that since experts have reported that there is discrimination in the test, the data must show it, although she said she didn't know how the data will help determine this. After writing her paper, however, she understood that the issue of discrimination is much more complex and that although her inquiry showed that there are some differences between Blacks and Whites, that the differences were not as great as she had anticipated: "Upon my investigation, I found that the differences that I expected to see where not as big as the many debates on the topic make them out to be" (p. 3).

This process of reflecting on her beliefs about discrimination and equity seemed to emerge not solely from the process of spending two weeks investigating data, but from the feedback these efforts had pushed her to articulate. For Chloe, a number of key elements during the inquiry process appeared to be critical for her deeper consideration of her topic. The *skills* and *background knowledge* she gained from the course in statistics, analysis of data with the software, and readings in equity, all in the context of accountability provided her with an authentic context, opportunity, and motivation to conduct her inquiry. Furthermore, having *time* to conduct the investigation allowed her to perform more than a superficial analysis of the question of study and allowed her to challenge her pre-conceptions that she would find evidence in the data of test discrimination. The technical *support* she received allowed her to get past initial frustrations and take advantage of the software in her analysis. Furthermore, *feedback* she received during her presentation caused her to reflect on subtle but important distinctions in equity concepts. The *validation* she received during the interviews, of her experiences and her ideas about minority discrimination, likely provided her with confidence she needed to further express her beliefs and articulate her emerging ideas

about discrimination and disparate impact in testing. Finally, multiple opportunities (*iteration*) and encouragement to *reflect* on and revise her ideas allowed her to deepen her understanding of very difficult concepts in equity and engage her further with the project and topic of her investigation. Chloe also expressed that her investigation was very personal to her, in that she was a member of both groups (female and Black) that she thought to be the result of discrimination.

From the presentations, papers, and focus interviews, it was clear that some students, like Chloe, gained a deeper understanding of issues of equity from their inquiry projects as well as being influenced by the results of their peers. For example, Chloe initially set out in her inquiry to use test results to "prove" discrimination, yet came away with an understanding that there were three separate issues to be considered when looking at differential performance: differences in outcomes, differences in opportunities, and difference due to bias in wording and context of questions. When probed in a focus interview, she noted that differences in opportunity and testing bias play out in differences one sees in outcomes. She used examples presented in the course and other teachers' presentations to illustrate differences in opportunity and funding. Her examples drew on results such as Rachel's presentation on the Robin Hood plan, José's project about differences funding in low-performing and exemplary schools, and Charmagne's claim that income was a tacit factor in SAT performance. Of the three kinds of differences, she said she thought that differences in opportunity and bias in testing were the ones that needed the most attention, rather than the focus on differential performance, which is what she thought at the beginning of the project.

Although Chloe had not previously taken any statistics coursework, she entered the course with one of the highest pretest scores in the class. Her pre-post tests did not indicate a strong improvement in her *understanding* of distribution and variation during the course, however she posted above-average increase in her *confidence* in understanding statistics. She indicated during her interviews that the issues about discrimination she learned from her inquiry and readings in this course coupled with her African-American Literature and World Culture coursework 'open your eyes' to these issues and she found the inquiry project very compelling.

	Posttest		Fathom Behavior		Statistical Use Level		Engagement	
Chloe (H, F)		Q4		Wonderer	√	5	√	High
		Q3	√	Wanderer		4		Moderate
	√	Q2		Answerer		3		Low
		Q1				2		

5.3.7 External Issues

Thirteen students' projects were described above. As there are only two projects remaining, these will be summarized briefly here. In both of these projects, conducted by Kathleen and by Rachel, the prospective teachers indicated only some personal connection that motivated their choice of a topic of inquiry.

Kathleen

An interest in studying drop out rates was partially motivated for Kathleen by a close friend who had dropped out of school in tenth grade. Kathleen's final paper described in detail the method of calculating drop out rate by the Texas Education Agency (TEA); her results consisted mostly of data tables (eighteen in all), for example showing drop out rates for a subset of states in the US in 1986 and 1997. She exhibited one set of her results as longitudinal line graphs comparing the drop out rates as calculated by TEA and by the Intercultural Development Research Association (IDRA), a watch-dog for minority rights. She also cited drop out rates, as a table and cumulative line graph, disaggregated by ethnicity and gender that were published by the National Center for Educational Statistics. Most of her data tables and graphs were copied as images off of the internet.

	Posttest		Fathom Behavior		Statistical Use Level		Engagement	
Kathleen (W, F)	√	Q4		Wonderer		5		High
		Q3		Wanderer		4	√	Moderate
		Q2	√	Answerer		3		Low
		Q1			√	2		

Rachel

Rachel an exhibited an interest in studying the Texas school funding legislation called Chapter 41 Wealth Equalization, commonly known as the Robin Hood Plan. She had no personal connection to the study except that her father had presented information about Robin Hood to the local community in his role on the board of trustees for her high school's educational foundation. For her study, Rachel chose a sample of five local high schools and investigated the longitudinal relationship between the school's overall passing rate on TAAS and whether they gave up or received funds under Robin Hood. She displayed her data as a series of scatter plots in Fathom with a line plotted that represented the least squares line summarizing the state passing rates from 1994 to 2002. For each graph, she made no mention of variability and concentrated only on two issues: whether the passing rates were higher or lower than the State ("above or below the line") and whether the district had received or given up funds. She concluded that high wealth districts, like her own, still performed above the State average despite the fact that they gave up funds.

She argued that the Robin Hood plan was equitable because it provided districts, regardless of tax base, with equal funds to educate their students. This belief mirrors that of equity as equality of inputs as described by Lynch (2000). She further argued that without the Robin Hood plan, people may believe that the difference in performance by districts was due to lack of funding. If people believe that, she claimed, then "there is no viable way to solve this issue of equity" (p. 13).

	Posttest		Fathom Behavior		Statistical Use Level		Engagement	
Rachel (W, F)		Q4		Wonderer		5		High
	√	Q3	√	Wanderer		4	√	Moderate
		Q2		Answerer	√	3		Low
		Q1				2		

5.3.8 Beliefs about Equity

The project descriptions above point to a number of surprising results (to me) with regard to equity: (1) before entering the course, many prospective teachers had neither encountered nor been asked to reflect on issues of equity, (2) these issues, although extremely contentious, were very difficult for the teachers to express and

openly debate, and (3) the beliefs that the teachers did express were extremely diverse. I believe that the beliefs documented here are likely incomplete and in fact some of the teachers showed that they hold multiple, possibly inconsistent beliefs consecutively or even simultaneously. Some of the beliefs that were documented align with those articulated by Secada (1994), Kahle (1996), and Lynch (2000). For example, beliefs about equity as evidenced by equality of inputs (e.g., resources) or equality of outputs (e.g., test scores), equity as concentrating resources for those who will benefit most (e.g. magnet programs), equity as uniqueness of individuals, and equity as fairness. In addition, several beliefs about equity emerged which are not aligned with these, such as equity as colorblindness (e.g., all races are equal, so ignoring race is the most equitable approach) and equity as consequential validity (one must determine what is equitable by examining the consequences of actions regarding subgroups).

5.4 ENGAGEMENT REVISITED

The choice of topic did not show appear to show any systematic differences in either understanding of equity or use of statistics, as evidenced by the sharp contrast in statistical use and engagement by those who chose similar topics. Eleven of the seventeen students who turned in projects chose topics to investigate that were similar to others in the class. If one were to match these students based on the topic of their inquiry, we can see that both the level of statistical use and the connection to equity is quite divergent. I conclude from this that the topic of inquiry did not in and of itself influence the level of statistical use or connection to equity.

Although the *topic* of inquiry did not appear to affect the level of statistical evidence used in their inquiry projects, their *engagement* with their topic did show some association to their level of statistical evidence used. I scored each of the prospective teachers on their level of engagement (Sections 5.3.1 - 5.3.7). It was already established (Section 5.3.1) that the level of statistical understanding as measured on the pre-post test showed no correlation with the level of statistical use on the inquiry projects. As might be expected, neither performance on the posttest and nor degree of improvement (pretest to posttest) correlated with engagement ($r = -0.04$, $r = -0.03$, respectively), although for two teachers in particular, Anne and Janet, it appeared that their interest in mathematics and statistics may have generated additional interest in the topic of equity investigated in their inquiry. Fathom behavior also did not show any measurable

association with level of engagement; the median level of engagement across the three behavior types was similar (Figure 5.15).

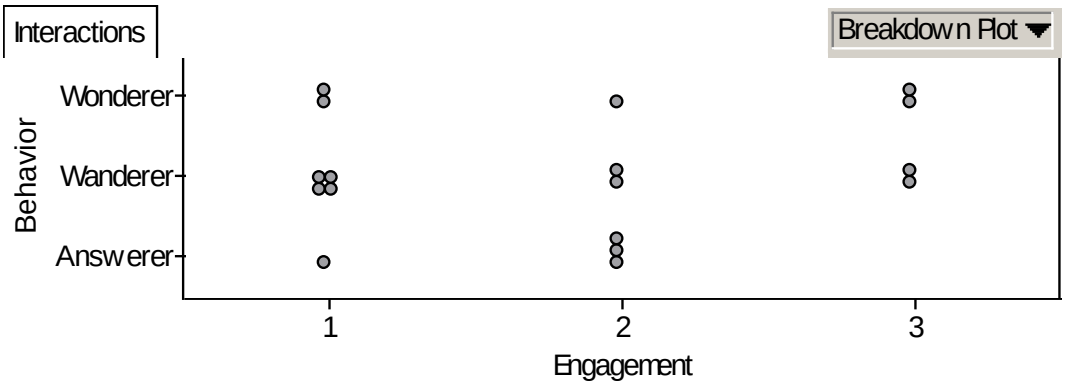


Figure 5.15: Breakdown plot of engagement level for each Fathom behavior.

It was conjectured that their engagement level might be a factor in their choice of statistics as a tool for evidence in their inquiry project and this relationship was tested. It was found that if engagement were converted from low-moderate-high to a numerical scale of 1-2-3, the level of statistical use showed a fairly strong positive correlation with level of engagement in the projects ($r = 0.72$). Figure 5.16 below shows that this was particularly true for those who registered at the extreme ends of the engagement scale. That is, everyone with a "high" level of engagement used at least concepts of variation and distribution (level 4 or above) as evidence in their projects whereas no one with a "low" level of engagement chose to use these concepts (level 3 or below). This figure also demonstrates a seemingly natural divide or threshold between levels three and four on the statistical use scale. This may indicate that concepts of distribution and variation are important notions that are qualitatively different conceptually than those included in typical descriptive statistics.

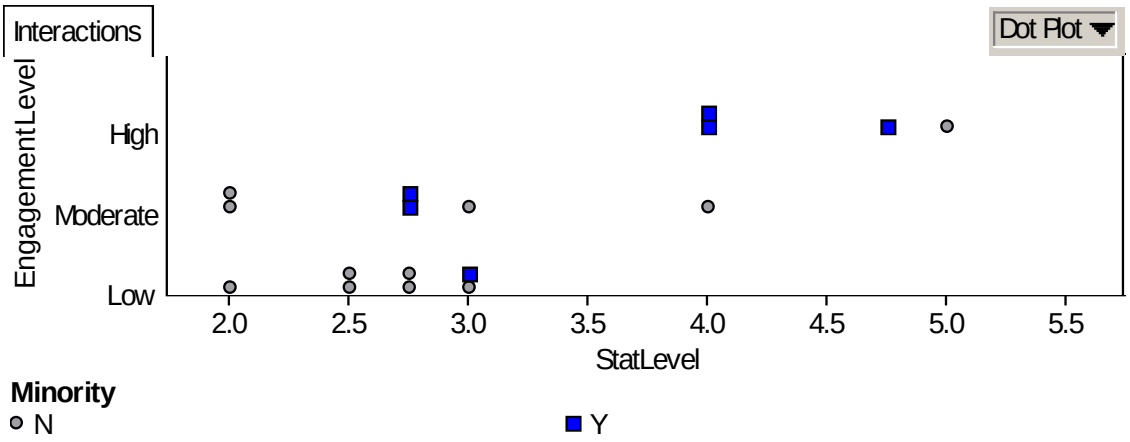


Figure 5.16: Relationship between level of engagement and level of statistical use. Minority students are shown in blue.

5.5 SUMMARY

What might be gleaned from the above cases is that for some teachers, regardless of previous statistical experience, the opportunity to dig into the data with a question that they found compelling and personally relevant may have led to stronger uses of both statistics and deeper explorations into equity. This is particularly poignant in the cases of the three minority women—Maria, Chloe, and Charmagne—who investigated very personal issues related to their own race, and used some of the highest levels of statistics in the class even though these same women had among the lowest scores on the posttest. The evidence provided by these three cases is an indication that for some preservice teachers, strong personal engagement with the equity topic of their inquiry can be a potent motivator for them to use more powerful statistical tools to provide evidence for their inquiry. In addition, other teachers, like Anne and Janet, were able to use their deep understanding of concepts of distribution and variation to develop strong inquiries into important issues of equity. Unfortunately, most of the prospective teachers in the course did not engage uses of statistics that involved variation and distribution. The prospective teachers that were less personally engaged did not make use of more powerful statistical concepts involving variation and distribution as tools for inquiry, relying instead on tables, summary statistics, and static displays in presenting their results. This was particularly true for those displaying the lowest level of engagement in their inquiry project. Unfortunately, the numbers in my sample are not

large enough, particularly among those who did choose to include more robust statistical concepts in their inquiry, to confirm that this is a systematic relationship, although the cases presented are powerful examples.

In addition, eight of the thirteen teachers who included a sample as part of their investigation used poorly-chosen samples in their final paper. This indicates that (1) as an instructor, this concept was not well-taught, and indeed in hindsight I can see I did not focus on this in class; and (2) these teachers likely did not see the potential of *statistical tendency* to make generalizations. The statistical level of the projects by these prospective teachers was on average more than one level lower ($\bar{x}$ = 2.7, s = 0.64, n = 8) than those who used well-chosen samples ($\bar{x}$ = 3.8, s = 0.75, n = 5), a significant difference between these two groups (p = 0.03), pointing to the possible importance of emphasizing proper sampling as a topic if there is a desire to move the prospective teachers towards a greater focus on variation and distribution in their use of statistical evidence in data-based inquiry.

My assumption, at the beginning of the study, was that if preservice teachers were provided with a relevant context to undertake in their inquiry, provided with background knowledge and the technological and statistical tools to investigate and analyze the data, given experiences with structured investigations that were modeled by the instructors, then they would seek to conduct interesting inquiries on a topic that was important to them. I further assumed that they would be able to structure the process, with some support, requiring them to (1) create a measurable question or conjecture, (2) find useful data and a valid sample among those used in class or available on the internet, (3) analyze the data with appropriate statistical evidence, (4) present a reasoned argument in their interpretation of their analyses that was related to their conjecture or question, (5) document additional questions that arose (and possibly following up on some of them), and (6) state conclusions from their investigation. In addition, I expected that they would be able to provide their reader with background information needed to understand the context of their study and link their investigation, through their readings, reflections, and discussions, to an issue of equity. In hindsight, I see that the process is probably more complex than I realized, with more potential links that could break down between finding a topic and communicating their results, although I don’t think my expectations were completely unreasonable.

There are many issues involved here: the complex process of an open-ended inquiry of an ill-structure problem, ability to apply statistical evidence in this kind of situation, engagement with their project, beliefs about equity, and issues of instructor support and previous inquiry experience of the prospective teachers. What I found is that for most of the teachers, a single experience in such a complex investigation, with good background information and content but perhaps insufficient time and support, was not enough for them to be able to go through the entire process at the level I had hoped. The purpose of the dissertation was to investigate as the primary question, the interaction between their understanding of variation and distribution and their ability to use it as evidence in articulating their understanding of equity in an environment designed to support this connection. My driving question assumed that this connection would exist, I had assumed I was only seeking evidence in how they chose to reveal the interaction between these two concepts.

The evidence found in this chapter documents that the experiences the teachers had in the course were successful in that they were able, in a supported setting with a well-structured problem, to "see variation" and recognize important elements of distribution when comparing groups. In addition, all of the teachers were able to conduct a semi-structured technology-based investigation with a large data set and displayed multiple approaches in the way that they used the technology to do this. They were comfortable with using the technology, they did not feel overwhelmed with thousands of pieces of information in the large set of data nor did they appear uncomfortable with the lack of direction as to *how* go about the process of stating a conjecture, seeking evidence to support their conjecture, and stating a relevant conclusion based on the evidence they found. By the end of the course, most of the teachers were still fairly uncomfortable, however, with openly discussing their most tacit beliefs about equity, but were still able to recognize and articulate problems in the current system with regards to equity, as evidenced in the debate about testing where they were assigned a position to argue.

Some of the results that were uncovered surprised me, for example that their decision whether to apply strong statistical evidence was *not* related to their level of understanding of statistics in a structured environment, once you assume that they had at least a minimal understanding of these topics. This is not to say that their learning of statistics was not important in their ability to apply statistical evidence to their inquiry.

Rather, that their understanding of statistics and experiences in the course were not a *sufficient* reason for them to conduct deeper investigations. Ideas on how this process might have been improved will be deferred to the final chapter. In addition, there was not a clear relationship between how they went about conducting a structured investigation and their ability to apply the same reasoning in an open-ended, ill-structured, complex, and potentially contentious problem. It is not surprising that the teachers' level of engagement would have an effect on the depth of investigation and type of evidence they would show, although I did not expect this to be the only factor that significantly contributed to the outcome. In considering the process and development of conducting ill-structured and complex inquiries of equity in high-stakes testing that our much more experienced research team engaged in at SYRCE over the process of three years, the difficulties encountered by the teachers in conducting similar inquires, while balancing so many new experiences, in just three weeks is more understandable.

The final chapter will further discuss the qualitative results from this chapter and the quantitative results from Chapter 4 with respect to the research questions, and broaden the discussion in how the results of these two chapters might be interpreted further. Potential changes that might have better facilitated the teachers' ability to conduct their inquiry of equity with stronger levels of statistical evidence will also be discussed. In addition, the final chapter will state the conclusions of the study, its limitations, and implications for research and practice.

Chapter 6: Conclusions

This is a crucial time in mathematics education, when schools and teachers are caught between two competing and sometimes contradictory demands: the mathematics reform as envisioned by the National Council of Teachers of Mathematics (NCTM, 2000), and the accountability system fueled by the *No Child Left Behind* directive to improve student performance, particularly for minority children, on state-mandated exams. A significant way to empower teachers to survive these competing pressures would be to increase their ability to interpret the statistical data by which they are being judged in the accountability system. This is in contrast to the current practice by many schools of hiring outside experts to interpret the data and then telling teachers what needs to be improved, based on their analysis. By teaching teachers the skills they need to conduct their own statistical inquiries and analyses of their students' data, they are simultaneously empowered to determine their needs according to the data and attend to requests by the NCTM and NSES Standards movement to improve their content knowledge in statistics and data analysis, experience with inquiry, proclivity towards equity, and facility with learning technologies. The study described here examines the interaction between teachers' statistical learning as they conduct analyses of student assessment data, and their inquiry into equity and accountability issues. This combination has potential to improve teacher learning in two critical areas (statistics and equity) while also empowering teacher professionalism at a time when many teachers are feeling "deskilled".

The study described in this dissertation examines preservice teachers as they use their learning and understanding of statistics, particularly concepts of variation and distribution, culminating in the conduct of a three-week inquiry into issues of equity through analysis of accountability data. The study used a mixed methodology to track their prior knowledge and its evolution from evidence collected as they conduct the inquiry, present their findings to their peers and instructors, and struggle to write the results. Particular attention is given to how the prospective teachers sought statistical evidence to support their emerging and at times emotionally charged theories about equity and how they used that evidence to make the case for the results of their investigation. In many cases, while their selection of inquiry topics revealed insight into the relevant questions, the use of statistical evidence was disappointing, despite

evidence of their understanding of statistics documented in other settings. Potential factors that may explain this are also examined.

This study brings together several elements that serve to strengthen and support one another: preservice teacher education, statistical content, equity, inquiry, accountability, and the use of dynamic technology. Independently, these are important areas of study. In combination, the potential is great for increased learning as well as efficient and powerful inclusion of these topics in an innovative teacher education program. I propose that the kinds of learning the teachers showed in the study were not only attributable to the inclusion of these topics, but the way in which each of these elements interacted in powerful ways to support, challenge, and motivate one another.

This closing chapter will first summarize the results presented in Chapters 4 and 5, then discuss what these results mean in relation to the central research question. Next, I will talk about the implications of the study for research and practice as well as its limitations. Finally, further research will be suggested to extend the results of the study.

6.1 SUMMARY

Many high-stakes test results are presented only as summary statistics (e.g. mean score or percent passing). Disaggregated, this can lead to stereotyping of students of color and those from families with fewer economic means. Concepts of variation and distribution in statistics require that one go beyond reducing subgroups to single-valued summary statistics. Rather, they require prospective teachers to see the whole as well as the parts and help them gain an appreciation of the diversity of individual students. In addition, they allow one to see segments of the population that are often hidden by the reporting of only measures of center and proportion passing—the lowest and highest performing students, for example. Results of the pilot study (Confrey & Makar, 2002; Makar & Confrey, in press) suggested that perhaps the experience of "seeing variation" (p. 10, Watkins, Schaeffer, & Cobb, 2003) may support and coincide with building awareness of the struggle of students who are often neglected, or acutely remediated, in the accountability system. With this in mind, the central research question that motivated this study was:

> In a preservice course created to support learning about assessment, technology-driven data analysis, equity, and inquiry, how do prospective teachers use the concepts of variation and distribution to support their understanding of issues of equity and fairness in testing?

This question implies that the study probe into the teachers' learning of statistical concepts, their articulation and understanding of equity, and use of technology to support their inquiry. The intersection of these elements was equally important. The main research question was supported by four sub-questions designed to assist in unpacking the larger research question and probe the intersections of these topics:

1. What level and types of understanding of the concepts of distribution and variation were learned? How did the teachers express this understanding in practice?
2. How was the technology used in relation to the students' inquiries? What behaviors did the prospective teachers exhibit in using the technology in a semi-structured investigation?
3. What can be said about preservice teachers' understanding of equity from their structured and ill-structured inquiry activities?
4. What is the interplay between the preservice teachers' statistical reasoning and the depth and breadth of self-designed inquiry into complex, ill-structured problems?

The subjects were eighteen university students at a large Texas university enrolled in a one-semester preservice course on assessment designed for prospective secondary mathematics and science teachers. The major artifacts collected during the course and analyzed in this dissertation include: a pre-post test on statistics, emphasizing concepts of variation and distribution; three reflection papers written during the course by the preservice teachers on various aspects of equity in the accountability system; video-taped final presentations and written reports of a three-week self-designed inquiry project conducted by the subjects; and interviews conducted at the beginning and end of the course probing the teacher's interpretation of test data as well as their technology use (only at the end of the course) in a semi-structured data investigation. Qualitative data (interview transcripts and written artifacts) were analyzed using Grounded Theory (Strauss & Corbin, 1998) while quantitative data (pre-post test results and categorical data from codes and scales created during qualitative analysis) involved a combination of visual documentation of relationships represented in statistical graphs with support by appropriate statistical tests.

The major results produced from quantitative analysis of the pre-post test showed three findings. Firstly, understanding of variation and distribution, for the group in general, significantly improved during treatment. Secondly, it was assumed that teachers who entered the study with some previous statistics would out-perform those who had not, but performance on the pretest was almost identical for these two groups. Those who had taken some statistics, however, significantly outperformed those who had not by the end of the study. Thirdly, although minority students entered the course with significantly lower levels of self-confidence in their understanding of statistical concepts, their confidence improved during treatment, with no significant difference between White and non-White students by the end of the course.

The quantitative results addressed the first research sub-question by examining the degree to which the prospective teachers learned concepts of distribution and variation. For example, recall that five categories of questions were tested on the pre-post test: histograms, variation, distribution, comparing groups, and properties of sampling distributions (Central Limit Theorem). Beyond the fact that the prospective teachers *learned* these concepts, patterns in performance showed that concepts of variation and distribution were the weakest areas on the pretest, but showed the highest gains on the posttest. The pre-post test also showed that this growth was particularly profound in problems that were contextual. That is, the teachers generally had good intuition on the pretest about variation in traditional school-like problems involving dice, but lacked intuition about the same concept set in a real-life contexts. The gap between these two settings was dramatically reduced on the posttest, in fact the strongly contextual problems about variation were among those that showed the greatest growth from pretest to posttest. The pre-post test also showed that on entering the course, the teachers had very little experience working with skewed distributions and data in histograms, for example, frequently ignoring the height of the bins in estimating the mean or median.

Five important results came out of the qualitative data. First, results from interviews indicated that the teachers were able to verbally articulate their attention to aspects of variation and distribution in the context of comparing test results, and this attention was greater after the course. The words they used to articulate “seeing variation” and distribution were more often informal. For example, in general they neglected to use formal concepts of standard deviation but frequently spoke of “spread

out" and "clumped" distributions. Secondly, the teachers exhibited three general types of behaviors in using the statistical software to support data investigations which I labeled as *Wonderers*, *Wanderers*, and *Answerers*. The Wonderers, which represented about one-quarter of the subjects (n = 5), used the software to test theories they had and to support "I wonder" questions from emerging theories that arose during the investigation. Half of the subjects (n = 8) were in the second group, the Wanderers, used the technology as a fishing net to look for theories based on results that "popped out" during their analysis. The third behavior group was the Answerers (n = 4), who used the software as an efficiency tool to find a single piece of evidence to test their theory.

The third important result produced from the qualitative analysis was found in the teachers' articulation of beliefs about equity in their projects, reflection papers, interviews, and class discussions. Six different beliefs about equity were recorded, equity as: equality of inputs, equality of outputs, individuality and uniqueness, color-blindness, consequential validity, and requiring that resources be concentrated on the most talented. A seventh approach to equity was one of avoidance of the issues.

The fourth and fifth important qualitative results arose from examining the depth of the teachers' statistical inquiries. The fourth result showed that the level of statistical evidence presented during their inquiries did not match the level of understanding of variation and distribution that they demonstrated during the pre-post test and interviews. Furthermore, there was no correlation between performance on the posttest and level of statistical evidence. This was contrary to what was conjectured at the beginning of the study. What did emerge, however, as a fifth major result, was that the level of personal engagement of the prospective teachers in their inquiry showed fairly strong correlation to the level of statistical evidence that they used. In particular, those minority students who investigated issues relating to their own race consistently produced the highest levels of statistical evidence in their inquiry despite the fact that these students showed among the lowest levels of performance on the posttest. Conversely, those who did not choose investigations of issues that were personal to them showed significantly lower levels of statistical evidence. In addition, students who took advantage of the statistical tools available to them were able to dig deeper into investigating their chosen equity topic.

The qualitative results addressed the remaining research sub-questions. The prospective teachers' use of informal terminology in comparing distributions during the

pre- and post-interviews demonstrated how the prospective teachers chose to articulate their understanding of variation and distribution in an applied context; so while the pre-post test showed that the teachers had learned these concepts, the interviews showed that they chose to articulate their understanding largely in non-standard language. The Fathom interviews, addressing the second research sub-question, documented behaviors that the prospective teachers displayed in conducting a semi-structured investigation using the software. All of the prospective teachers demonstrated facility with conducting a data investigation in the software with a fairly large data set (fourteen variables and nearly 300 cases). They articulated a reasonable conjecture, supported it with evidence, and stated a relevant conclusion. However, few of the teachers took advantage of the ability of the software to display distributions or investigate variation in the results of their inquiries, and there was no obvious association between the behaviors the prospective teachers showed in using the software in a semi-structured investigation in the Fathom interviews and the likelihood that they took advantage of the software in seeking statistical evidence (beyond reporting means or percentages) in their inquiry projects. The third research sub-question was developed to seek insight into the teachers' beliefs about equity and the qualitative results documented a wide range of these beliefs. The final sub-question, investigating the interplay between their statistical reasoning and the depth of their inquiry of an ill-structured problem showed that there was no correlation between their *understanding* of variation and distribution, as measured by the posttest, and their ability or choice to *apply* these concepts, once they had gained a basic understanding from their experiences in the course. This final result was contrary to what was conjectured at the beginning of the study, where it was assumed that understanding of variation and distribution would be strongly correlated to the prospective teachers' depth of inquiry of equity.

The results summarized above addressed the research sub-questions posed in the study, however these results could be interpreted in multiple ways. My interpretation is provided in the discussion below.

6.2 DISCUSSION

The results presented in Chapters 4 and 5 provide insight into what statistical concepts the prospective teachers learned and were able to articulate, their use of technology, beliefs about equity, and evidence they sought to present the findings from their inquiries. It should be noted that the purpose of the dissertation was not to study

these elements individually, but to examine what relationships might exist between understanding of distribution and variation, equity in the accountability system, and inquiry using high-stakes test data with technology. In addition, the purpose was not to evaluate the effectiveness of the treatment (the preservice course).

The setting of this study in the preservice classroom was strategic in that many of the experiences the prospective teachers encountered were new to them. Teacher education is critically in need of providing teachers with patterns of thinking and habits of mind before they hit the overwhelming demands of juggling a classroom for the first time. The statistical component of the study has the potential benefit of improving not only their content knowledge in a domain that is increasingly becoming central to the secondary mathematics and science curriculum, but other likely benefits as well. For instance, I would hope that their experiences engaging in an independent inquiry themselves would increase their disposition to support their own students' inquiry activities. As discussed below, the statistics that they used provided teachers opportunities and tools to dig deeper into the inquiry of issues that impact not only their students, but the teachers' tacit stereotypes and very personal beliefs about their students.

6.2.1 Statistical concepts of variation and distribution

The prospective teachers' understanding of variation and distribution demonstrated different dimensions within each form of data collection, including demonstration on the pre-post test, articulation in the interviews, and application in the final inquiry projects. In addition, confidence issues about these topics surfaced in the pre-post test. This is the first point of discussion.

Confidence

The results of the pretest indicated that there was no significant difference between the understanding of statistics by minority students and by students of the majority culture. Of great concern, however, is that the minority students in the course entered with a significantly lower level of *confidence* in their understanding of statistics. Although this result was unanticipated, I begin with the discussion of it because I believe it to be important. Other research has pointed to confidence as a key reason that many minority students do not choose to study higher level mathematics and science (Seymour & Hewitt, 1997). All of the subjects in the study were already mathematics or

science majors, so the fact that these minority students still entered with low levels of confidence in their understanding is a concern, since these are students who likely had higher levels of confidence than their peers who were not math or science majors. It is very encouraging that through the experience of the course, the minority students were able to increase their level of confidence to narrow the gap with their White peers. Although the study indicates that their confidence improved during the course, it is not known what aspect or aspects of the course influenced this improvement. It is presumed, but not confirmed by this study, that this increase in confidence was a result of the teachers learning statistical concepts in a context that was compelling to them and that the statistics became a tool of inquiry into issues of equity for many of them. Furthermore, a focus on equity in the course likely gave legitimacy and voice to their experiences and identity as students of color.

Seeing variation

Variation and distribution are critical to the understanding of statistics (Wild & Pfannkuch, 1999). The prospective teachers demonstrated that they were attentive to aspects of the distribution and were seeing variation during interviews that asked them to compare distributions. The results of the pre-post test corroborate this result. They spoke informally, but their language was rich. This can be taken as a consequence of the approach of the treatment, under the assumptions of the study, that emphasized informal understanding in preference to or along with development of formal statistical concepts. Together, these results indicate that in a structured setting the prospective teachers were able to pay attention to important aspects of variation and distribution. Work in statistics education indicates that both children (Konold & Higgins, 2002) and inexperienced adults (Confrey & Makar, 2002) often begin their work with data focusing on single data points or summary statistics. If we want students to move from this focus on individual points towards a distributional perspective of data it is critical for teachers to have this perspective themselves. The results of this study are promising in that all of the teachers by the end of the study were able to articulate that they were looking beyond single points and summary statistics when describing distributions. This was after only a relatively short (one-semester) preservice course that was focused not on statistics itself, but being able to *use* statistics as a powerful tool to interpret data.

The emphasis in learning statistics for university students and preservice teachers is often focused on either descriptive statistics or theoretical probability

distributions and hypothesis testing. I want to make a particular point here about the language the teachers used when they described distributions, because I think it has important implications for both research and practice. Research has been developing in the last few years about middle school students' conceptions of variation and distribution through the work of researchers such as Paul Cobb, Cliff Konold, Arthur Bakker, and others. Through their work, the research community has shown it values less formal language, like "spread", as valid forms of statistical conceptual thinking in students. Contrast this with the definition of spread in the *Cambridge Dictionary of Statistics* (Everitt, 1998). Here, the word *spread* is denoted as a "synonym for dispersion" (p. 316), which has the following as its definition:

> **Dispersion:** The amount by which a set of observations deviate from the mean. When the values of a set of observations are close to their mean, the dispersion is less than when they are spread out more broadly. See also variance. (p. 104)

This definition hardly captures the kind of thinking that the preservice teachers in the study articulated when they used the word *spread*, nor do I think it encompasses the richness of the concept of variation that the field intends. Statistics has been called the study of variation, yet as a discipline, we have no definition of what variation is (Reading & Shaughnessy, in press). In many ways this is fortunate, as it could have been given a definition far too narrow at a time when little research or curricular interest had been focused on concepts of variation. Ten years ago, concepts such as Bakker's (2001) "bump" or Konold et al.'s (2002) "modal clump" would likely not have been even considered statistical concepts. These terms, however, do capture an essence of "seeing variation"—undefined, yet primary to statistical reasoning—that more formal terminology and procedures often do not. By ignoring nonstandard terminology, we are neglecting rich sources of information about understanding of concepts of variation and spread.

Recently, children's usage of these less formal terms has received more attention, however, it is pretty much assumed that by adulthood, one should use "proper" terminology. Three major benefits come out of teachers' use of their own, informal terminology in describing, interpreting, and comparing distributions. For one, they are using words that hold meaning for them and that convey their own conceptions of variation and distribution. In the constructivist perspective, knowledge is not conveyed through language but must be abstracted through experience. Informal

language carries a subjective flavor that reminds us of this. Through interaction with others, this subjectivity becomes *intersubjective*—one's meaning is not identical to another's, but through further explanation, our meanings become more compatible with the language of our peers (von Glasersfeld, 1991).

Secondly, informal uses of language are more accessible to a wide variety of students, allowing entrance to these difficult concepts while encouraging teachers to be more sensitive to hearing rich conceptions of variation and distribution in students' voice (Confrey, 1998), words that may allow for easier access for students to class discussions. This is a more equitable, more inclusive stance; one that is contrary to the conception of mathematics (or statistics) as a gatekeeper.

Thirdly, if the goal is to provide students with experiences that will provide them with a more distribution-oriented view of data, then less formal language can help to orient students (and their teachers!) towards this perspective. Describing a distribution as "more clumped in the center" conveys a more distribution-oriented perspective than using, say, standard deviation or range to compare its dispersion.

Even if there is disagreement about the value of informal terminology in prospective teachers who are learning statistics, the teachers clearly had a fairly strong distribution-view of the data. After a relatively short period of instruction, virtually all of the teachers were able to provide a view of the data that goes beyond the black-and-white dichotomy that is frequently seen in students who study statistics formally (Abelson, 1995; Gardner & Hudson, 1999; Reichardt & Gollob, 1997).

6.2.2 Beliefs about Equity

The prospective teachers exhibited many of the same beliefs about equity as those articulated by Secada (1994). Equity as equality (of inputs or of outputs) was among the most common belief cited. For example, Emily focused on equity as equality of outputs when she expressed that there was no racial discrimination in the way that McCallum treated its African-American students (requiring they attend lunchtime tutoring) because in the end, their scores improved. However, José, as well as Angela and Gabriela, argued that because schools were all different, it was inequitable to evaluate and attend to them equally. Others, such as Brian, expressed a belief in equity as one where resources were used in ways that were most beneficial to the whole (biggest bang for the buck). Emily and Mark, in their examination of minority performance in Texas, took an equally pragmatic stance—that helping minorities made

sense because they would soon become the majority race in Texas and the economy would suffer if their level of education did not improve.

Other beliefs emerged that did not coincide with Secada's framework, for example beliefs about equity as fairness with respect to the *consequential validity* of the system, that is the intended and unintended consequences of the interpretation of test scores in the short and long term (Messick, 1995). Chloe questioned, for example, whether the test was really being used for an appropriate purpose, and worried about the results of decisions that might be made based on the scores. Angela and Gabriela were concerned about the validity of the method of evaluating schools and how this might impact schools and students. Christine and Anne each questioned the validity in ways that the accountability system might unintentionally discriminate against schools with small subgroups or encourage schools to undertake questionable strategies to improve their ratings. Concern about fairness regarding the consequential validity resulting from the accountability system was also expressed by Charmagne, in her concern that the SAT's may unfairly hinder low-income students being accepted into college.

Discussions of race created a great deal of tension during the course. Several students, all White, objected to these discussions as they felt that they set back the clock on racial equality. Many indicated a belief that society, and teachers in particular, should be race-blind. The danger of this belief is that being blind to issues of race presumes that racial issues either do not exist or are not in need of attention. The minority students, however, expressed otherwise. They acknowledged that there were problems in how minorities were treated and argued that these problems should not be ignored, but attended to. Because of many of the prospective teachers' discomfort with articulating their beliefs about equity, it was possible that their demonstration of statistical competence in their inquiry projects was lessened or restricted.

6.2.3 Statistical Inquiry

Admittedly, I was disappointed in the overall level of statistical evidence the prospective teachers used in conducting their inquiry, particularly given that they were able to demonstrate much higher levels of understanding from the post-test and interviews. Several factors may have contributed to this. Firstly, the final projects were worth a considerable portion of their grade and they may have lacked confidence in their ability to use the statistical concepts correctly that they learned in such a high-stakes situation. Their use of informal language in the interviews may also point to an

underlying conception of variation and distribution, but insufficient formal understanding to apply these concepts. Another possibility is that the discomfort they experienced in conducting their inquiries left them at a loss as to how to make use of statistical evidence. The examples that were discussed in the course were pre-selected by the instructors and hid much of the hard work and three years of inquiry that the instructors themselves had done in order to uncover these exemplars. When the preservice teachers were unable to produce evidence as clean as what they had seen in the course, they may have lost confidence in their ability to apply statistical concepts to their analysis.

Statistical inquiry requires a great deal of time, reflection and support. It is likely that the prospective teachers needed to begin their inquiry projects sooner so that the difficulties they experienced in deciding on a topic and finding data did not eat into the time they needed to look for stronger evidence to back their claims. That so many of the teachers had never conducted an inquiry with such a complex and ill-structured problem indicates that this is a short-coming in their experiences both in school and in university. It is likely that a teacher who lacks confidence and experience in open-ended inquiry will not engage his or her students in this sort of practice. Inquiry is strongly advocated by both the mathematics and science education communities (National Research Council, 1996; 2000; 2001a), so this is of great concern.

Those who did use concepts that integrated variation and distribution with their inquiry were few, only five of the seventeen students who completed their inquiry projects produced levels of evidence beyond static displays. These five cases are of interest. Three of these were minority women who conducted inquiries into very personal issues of race and class. The other two students, both post-graduates, had very strong backgrounds in statistics to begin with and were able to use their understanding to create very interesting investigations. The depth of their inquiry was not just in the statistical evidence they used but also in very critical issues of equity – economic factors contributing to performance gaps between urban and suburban districts (Janet) and the effect of various actions taken by schools on their risk of being low-performing (Anne).

The prospective teachers' final presentations and papers demonstrate the difficulty they had in bringing together the concepts of inquiry, data analysis, and equity. In their projects many students regressed to using familiar tools such as

percentages and data tables, although in testing and interviewing in structured settings, they attended to variation and distribution. This suggests the importance of engaging in open-ended investigations multiple times in order to develop a more sophisticated ability to integrate across these skills in a less structured inquiry.

6.3 LIMITATIONS

This study is subject to several limitations. Although the pilot study worked with practicing teachers, this study focused on preservice teachers; it cannot be assumed that the results presented here would transfer to work with practicing teachers. Results of the pilot study (Confrey & Makar, 2002; Makar & Confrey, in press), and Appendix A should be consulted to attend to results found with this population. Another limitation of the study is that it was not designed to study teachers' statistical reasoning nor their beliefs about equity systematically. Although some results address these topics, they are not comprehensive enough to stand alone. Several of the results could be interpreted as meaning that the particular treatment may have lead to an increased in understanding of statistics and issues of equity. However it should be cautioned that the research here was not designed to study the effectiveness of the course. This would require an evaluation study with comparison groups and different measures in place to capture ways in which the course *changed* teachers' beliefs about equity and proclivity with inquiry, for example.

6.4 IMPLICATIONS

Implications for this study can be located both for the research community and those involved in the practice of teaching teachers.

6.4.1 Implications for Research

The topics contained in this study provide the greatest feedback for the research community as an integrated whole, particularly those interested in research in teacher education. However, I have broken down these implications for researchers in statistical reasoning and equity separately. Other areas of research can also benefit (e.g. technology use, inquiry), but because the results presented here focus on the integration of statistical reasoning and equity, these will be the focus here. Later publications are planned to discuss further the issue of inquiry in more depth.

Research in statistical reasoning

Recently, there has been an increase in interest among the statistics education community in the conceptual understanding of variation and distribution in children. Researchers have paid particular attention to ways in which students used informal language to express their understanding of these concepts. What has not been studied is how adults may also use informal language to express their understanding of variation and distribution. The study presented in this dissertation presents a possible new direction of thinking about prospective teachers' conceptions of variation and data distributions. This study has shown that secondary preservice teachers, majoring in math and science, use informal terms; other research (e.g. Bakker, 2001; Konold et al., 2002) has shown that students do as well. It has been suggested by a statistician (Smith, 2004), but not yet researched, that professional statisticians also use informal language, such as talking about the *shoulders* and *tails* of a distribution. This kind of non-standard, informal use of language use needs to be given a greater emphasis in research on statistical reasoning.

Research on adults' statistical reasoning has often focused on descriptive statistics (e.g. graphical interpretation, measures of center and sometimes spread), or inferential statistics (e.g. sampling distributions, hypothesis testing). The latter two are often the only two types of statistical training offered for teachers. I would argue that an intermediate level of understanding, located between descriptive statistics and inferential statistics, is needed to promote broader awareness of concepts of distribution and variation. Although descriptive statistics and graphs are important, training in this area doesn't provide insight into the power of inferential statistics; and typical coursework in inferential statistics is frequently too focused on formal hypothesis testing and can neglect the more subtle and less theoretical aspects of distributions. Given the ability of software packages like Fathom to create graphs easily and highlight selected data in multiple representations, there is an opportunity to provide adults with an understanding of *statistical tendency* and informal concepts of inference without the necessitating the introduction of formal hypothesis testing. A focus on statistical tendency provides an opportunity to bring concepts of variation and distribution into the level of statistical literacy that extends beyond simple displays and summary statistics. Further research into this middle ground is necessary to better understand the opportunities, limitations, potential drawbacks, and approaches that could facilitate

conceptualization of statistical tendency in adults. Some research has begun in this area in the last few years (Biehler, 1997, 2001; Makar & Confrey, under review; Rubin, 2002), particularly among those involved in the International Research Forums on Statistical Reasoning, Thinking, and Literacy (SRTL), but it needs broader emphasis.

Research in equity

This study implies that the preservice teachers' beliefs about equity were diverse. Many of them, particularly those who were White, held the belief that the most equitable thing for them to do is to be race-blind. Others advocated for school-based programs that increased student scores, because "they worked", regardless of how minority students were treated by the process of remediation or the potential consequences for individual students. This is of concern because in many preservice programs, these beliefs are not openly discussed. This study brings to light the need to continue to focus on preservice teachers' beliefs about equity, particularly in an environment where both White and teachers of color can interact. For example, many of the White students objected to the use of the term "children of color" until one of the minority students finally made a comment that she was not offended at all by this term. This interaction was critical for many of the White students to be able to move towards a more empathetic stance towards equity, rather than believe that the issue was best ignored. Furthermore, it gave the African-American and Hispanic preservice teachers an opportunity for their beliefs to be legitimized.

The difficulty that the prospective teachers had in reflecting and articulating their own beliefs is in contrast to their ability, by the end of the preservice course, to argue both sides of the issues in a debate about using high-stakes testing for critical decisions such as graduation, promotion, and tracking as well as its use for special populations, like English-language learners and students with learning disabilities. In the latter environment, the teachers provided coherent, well-supported, and articulate arguments and were able to both accept critique and challenge the beliefs of their peers. The important difference is that in the debate, it was unknown to the audience whether the beliefs they were arguing were their own or the opposing viewpoint because the position they were arguing was randomly assigned to them a few minutes before the debate. What this indicates for research is that there needs to be a better understanding of the elements needed to assist teachers in informing, reflecting on, articulating, and

challenging beliefs about equity in education. This was an important result of this study, but needs further research.

6.4.2 Implications for Practice

Implications for practice will focus on teacher educators, both in professional development of practicing teachers and schools of education that prepare preservice teachers. Further implications, for those who teach statistics, will be directed towards both teacher educators and teachers of statistics. Implications with regard to inquiry will focus on both teachers and those who teach teachers.

Professional Development Practices

Cohen and Hill (2001) noted in their large-scale study of California teachers that one of the only factors that correlated with reform-based teaching practices and beliefs was professional development experiences that centered around *content* and *student assessment*. The results of this study point to a promising parallel experience for preservice teachers providing them both an opportunity to improve their statistical content knowledge and work to better understand student assessment. This study provided an example of an environment where preservice teachers (and presumably practicing teachers) can build a deeper understanding of statistical concepts in a compelling context, while at the same time broaden their perspective of equity and deepen their understanding of the benefits and limitations of large-scale testing. Schools and teacher educators can modify the course described here and elsewhere (Confrey et al., 2004) to meet their own needs. Furthermore, it is hoped that the integration of important topics in teacher education used in this course (inquiry, equity, content knowledge, technology use) can serve as an example and motivator for program directors to further integrate other teacher education courses to attend to the problem of disconnected coursework in preservice programs (National Research Council, 2001a).

Teaching statistical reasoning

Given the increased emphasis on testing that narrows statistical reasoning to traditional measures of center and basic reading of tables and charts, there is concern that teachers may not provide their students with statistical experiences beyond this narrow perspective. If there is a desire to move students—and given this study, teachers as well—towards more informal language and rich, meaningful experiences with data,

then teachers need to be given the opportunity to learn statistics in a compelling environment, where the focus is not on methods and procedures but concepts and tools for inquiry. In addition, teachers need to develop understanding and respect for the informal language their students use when describing distributions. There are several reasons for this. For one, teachers need to learn to recognize and value informal language about concepts of variation and spread to better attend to the ways in which their students use this same language. Secondly, although the teachers in this study are using informal language, the concepts they are discussing are far from simplistic and need to be acknowledged and valued as statistical concepts. Thirdly, encouraging the use informal terminology may help teachers and students to better switch to more formal terms when (and if) appropriate; the scaffolding of these more informal terms may then help to redirect students away from a procedural understanding of statistics and towards a stronger conceptual understanding of variation and distribution, as well as to provide teachers better insight into students' understanding. A fourth benefit of using informal language is to increase students' access to statistics. Because science and mathematics are often held up by society as requiring special talent in order to be able to pursue these subjects successfully (Toulmin, 2001), informal language has the further benefit of breaking down the "mystique" of the subject (Lemke, 1990), which has been shown in science and likely also exists for statistics. Valuing informal language statistics allows students and teachers an opportunity to express their own meaning-making rather than attempt to use formal language that may not yet hold meaning for them, or worse decide that they cannot "do" science. This can work to make the subject more inclusive, particularly for those who are traditionally under-represented in the math and sciences.

Many of the prospective teachers in the study had already had some statistical methods either in a statistics course or else as part of their mathematics and science coursework. It is of concern, given the results of the pretest, that they showed no better performance than those who had not previously learned statistics. This implies that one cannot assume that prospective teachers will develop an intuition about variation and distribution in their regular statistics coursework. The experiences the preservice teachers had in this study with analysis, as well as the compelling context, likely contributed to their increase in understanding of variation and distribution even over those with no previous statistical experience.

Teaching and Learning Inquiry

This study points to the lack of opportunities even majors in mathematics and science have with conducting inquiry. This implies that there is a great need to increase and support teachers' experiences with conducting inquiries of complex, ill-structured problems. If teachers are not comfortable themselves with inquiry, it is unlikely that they will provide their students with similar experiences. Thus, if the vision of inquiry put forth by the National Science Education Standards (National Research Council, 1996; 2000) is to be realized, teacher education programs and professional development programs must consider providing teachers to study and experience inquiry. The findings of this study indicate that while the preservice teachers were able to recognize variation and distribution, most did not choose to use these concepts in their own inquiry. This implies that teacher educators need to provide prospective teachers with multiple experiences conducting inquiry of complex, ill-structured, and authentic problems where they are given sufficient time, relevant and thoughtful feedback, opportunity for reflection, and ample technical support and validation. The results of this study imply that one cannot assume that one such experience will be sufficient for them to feel comfortable.

6.5 FURTHER RESEARCH

This study points to a number of areas that could extend the work done here. In statistical reasoning, for example, it would benefit the statistics education community, as well as teacher educators, to have a more complete picture of the understanding preservice teachers, particularly those in secondary math and science, have of statistical concepts beyond simple descriptive statistics and graphs. This population will likely have an increased burden in coming years of instilling school children with an understanding of statistical concepts. It would be of use to know to what extent teachers are prepared to undertake this burden. The results of this study provides a new look at examining teachers' conceptions and terminology in data analysis, but additional research is needed to determine more about how teachers understand concepts of variation and distribution. We also need to document better how teachers support their students' emerging statistical understanding.

When students and adults learn and use descriptive statistics, to what extent do they condense the distribution to this single value? That is, does the mean represent a

center or is it conceived of as representing the distribution as a whole? In the context of analysis of test data, for example, it would be of interest to know, to what extent do single measures (mean score or percent passing)—without regard to distribution, variation, and longitudinal trends—exasperate stereotypes?

Little research has been done on the use of statistics and inquiry in complex settings or with ill-structured problems. This is a much needed area of research as the types of problems in which statistics can become a powerful force are problems that are not well-defined and likely involve a great deal of complexity. In addition, a better understanding of the influences of iteration, feedback, support, validation, time, and opportunities for reflection are needed. The study reported here is only a beginning.

The second research sub-question probed into ways that the prospective teachers would use technology to investigate test data in support of a conjecture they made. Three behaviors were documented by the study; these behaviors can begin to help us understand both the benefits and drawbacks of using technology in exploratory data analysis. Although the behaviors exhibited did not produce any significant differences either in understanding of statistics or engagement in inquiry, researchers can use these behaviors to conduct further research into potential influences these behaviors exhibit on other aspects of technology use, or with a study similar to this with more focused attention to this aspect.

In the area of equity, the field would benefit from a more comprehensive study on preservice teachers' conceptions of equity. In addition, it would be useful to know what kinds of experiences these teachers, including and in addition to experiences described in this study, to move prospective teachers towards a greater understanding and caring about equity.

6.6 CONCLUDING REMARKS: RETHINKING TEACHER EDUCATION

The study documented in this dissertation provides insight into the potential of creating a more cohesive preservice teacher education program for secondary mathematics and science teachers that addresses and integrates issues of equity, provides authentic experiences in inquiry-based learning, improves teacher content knowledge in statistics and facility with dynamic learning technologies, and deepens understanding of the accountability system. This integration of topics provides benefits beyond creating an efficient way of including several important topics in teacher education or even giving the teachers a more authentic experience. The combination of

these topics in a relevant and compelling context can actually support improved learning in each of the topic areas above what might have occurred had they been taught separately. What is also documented here is that these experiences in a single course are not sufficient. Instead, these are topics that need to be integrated within the program in such as way as to provide prospective teachers multiple opportunities to reflect on their beliefs about equity and conduct ill-structured inquiries in different settings.

A very important subgroup of prospective teachers was served by such a strong focus on equity in the course; schools are terribly short of minority teachers and the opportunities to investigate issues of equity increased many of the minority students' interest and engagement as well as their ability to use their statistical knowledge to provide compelling evidence of their inquiry. The difficulties experienced by many of the prospective teachers in this study point to a problem not with the teachers, but with the approach to preparing them that lacks coherence and authenticity. Greater focus on experiences like the one described in this study is critical to empowering teachers with the tools they need to create more equitable learning environments for their students.

Appendix A: Pilot Study Description and Results[1]

The Systemic Research Collaborative for Education in Mathematics, Science and Technology (SYRCE) at the University of Texas at Austin conducted a series of studies of methods of research partnership with schools in order to undertake effective reform, through an approach called implementation research (Confrey, Castro, and Wilhelm, 2000; Confrey, Bell, and Carrejo, 2001) and in which the SYRCE model of systemic reform (Figure 1) was investigated. The SYRCE model (Confrey, Bell, and Carrejo, 2001) is a systemic perspective that contends that the professional development of teachers, based on the analysis of student assessment artifacts and data, will ideally improve teachers' content knowledge and develop their sense of community as learners, which will lead to better instruction with technology, which in turn will positively impact student outcomes.

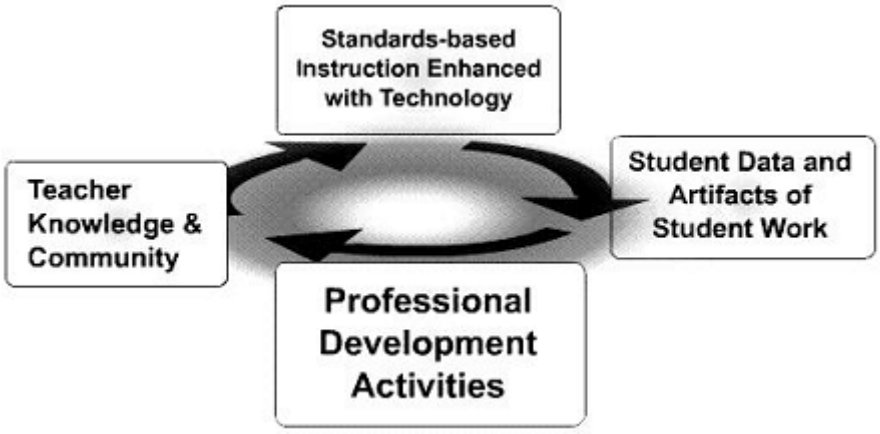


Figure 1: SYRCE model of implementation research in systemic reform

One of the projects undertaken by SYRCE focused on the professional development of teachers through the use of student assessment data (Confrey & Makar, 2001a; Confrey & Makar, 2001b; Confrey & Makar, 2002; Makar & Confrey, 2001; Makar & Confrey, 2002; Makar & Confrey, in press). The project was conceived as a mathematical parallel of the Writers Workshop from the National Writing Project (2002a, 2002b), where teachers learn to write rather than to be teachers of writing, and had a set of five related objectives:

[1] Much of the descriptive material here was adapted from Confrey & Makar (2002) and Makar & Confrey (in press) to provide background to committee members who are unfamiliar with the project.

(1) Strengthen *teacher content knowledge in statistics* by giving them the opportunity to learn statistics well beyond their curriculum, rather than learn how to teach statistics
(2) Immerse teachers in *focused investigations and chains of reasoning* about student performance data at an urgent time in a high-stakes accountability environment
(3) Build teacher *confidence and facility in using dynamic statistical software* (*Fathom*)
(4) Orient teachers with a *healthy mindset about data and inquiry*: the acceptance of uncertainty when searching for solutions, and the limitations and misuses of statistics and inferential reasoning
(5) Provide teachers with an opportunity to be *learners* in an environment that models Standards-based teaching

The long-term conjecture for the project was that when teachers are immersed in content beyond their curriculum in a context they find compelling and useful, this experience would transfer into improved classroom practice. Specifically, teachers will teach statistics more authentically if their understanding of statistics and how it can be used is developed through their own investigations as statisticians. Another longer-term conjecture was that as a result of these experiences, we should see teachers more willing and able to use technology in their classes, particularly when teaching statistics. These conjectures have not yet been tested.

Texas has a high-stakes accountability system. Students are tested in grades 3-8 and 10 on the Texas Assessment of Academic Skills (TAAS); high school graduation depends on passing in grade 10. Schools and teachers are held accountable for their students' performances. In this climate, urban schools that serve less academically advantaged children are under increased scrutiny and pressure to ensure they do not receive unacceptable ratings. As a result, professional development time in the schools is typically spent focusing on their TAAS results. These results are shared with teachers and schools in the form of summary data and hardcopies of individual student performance. Teachers feel that the accountability system creates a context in which they are under the gun and over which they have very little power. Their stance is reactive, rather than proactive. This context seemed ripe to invite teachers to examine the statistical data as investigators.

The teachers in the project initiated from the mathematics department at our partner school, an urban middle school that feeds into a low-performing high school. The department had five Hispanic women, one African-American woman, and one white male with experience ranging from none to over twenty-five years (median 8 years). Only three of the teachers were certified in secondary mathematics. The demographics of the school were 72% Hispanic, 21% white, and 7% African American. Sixty-two percent of the students were classified as Economically Disadvantaged.

The research conjecture was measured with a pre-post test of statistical content. Additionally, classroom observations were made before teachers began the study providing a baseline of their teaching practice of statistics and mathematics. At the end of the summer institute, teachers presented their findings to their colleagues and an audience of researchers. Clinical interviews at the conclusion of Phase II provided important triangulation of the data collected from the session videos, observations, pre-post tests, and videoed final presentations.

The implementation of the project consists of two phases, which occurred over the course of 6 months. Phase I was carried out from January to May 2001 with two full day and three 2-hour after-school sessions immersing teachers in exploratory data analysis, examination of accountability data, and time for orientation of the dynamic statistical software, *Fathom*. Approximately half of the contact time during the project was spent in front of the computer: learning the software, mirroring hands-on activities, creating simulations, testing conjectures, and searching for and investigating data. The software that we chose to use, *Fathom*™ (Finzer, 2001), is unique in its application as a teaching and inquiry tool. Whereas most statistical software tends to be like a "black box" (data in, answers out), or designed for very specific kinds of tasks, *Fathom* can be used to investigate a broad range of tasks at both an elementary and intermediate level. In addition, many schools in the district had already been purchased *Fathom* (although it was not yet widely used). During sessions, teachers worked to create increasingly more robust formulations and investigations of conjectures about their students' data while simultaneously building on their statistical thinking, reasoning, and content knowledge. Phase II of the project, a two-week summer institute, provided teachers time and support to further probe specific areas of inquiry of their own choice, examine deeper statistical concepts and tools as they were needed, and present their findings to their colleagues and a group of researchers. Phase I was repeated, by request, as a stand-

alone professional development workshop for two intensive days just prior to the beginning of Phase II and included nine secondary math and science teachers from surrounding schools. A high school math teacher and preservice math teacher stayed on for Phase II; only two teachers in the partner school were able and obligated to attend the summer institute (due in part to a change in principal at the school).

Typically work sessions began with some limited exploration of a data set as an impetus for continuing with training in the use of the software. A hands-on activity of a statistical concept followed, which was then continued on the computer with the assistance of the newly acquired skill in the software. Teachers were initially introduced to and explored the ideas of central tendency, distribution, variation (particularly as it related to small sample size) and graphical representations of data: boxplots, histograms, dotplots, and scatterplots. During the summer institute, statistical concepts were expanded to include correlation and least squares regression, influential points and outliers, standard deviation and variance, the central limit theorem, confidence intervals, sampling distributions, the null hypothesis, z-scores, and t-tests. We delayed introducing statistical tests until nearly the end of the summer institute, as we feared that the premature use of significance tests would aggravate a mindset of the accept-reject dichotomy of statistical tests (Reichardt & Gollob, 1997; Abelson, 1995). Sampling distributions were used frequently in class activities to instill a tolerance for variation and to provide a conceptual foundation for confidence intervals, p-values, and t-tests without focusing on formulas and rules. Software instruction was provided on graphs and statistical summaries, importing data, least squares regression, relational graphing, sampling, simulations, hypothesis testing, and more advanced software features (collecting measures, scrambling, and stacking). At an informal level, teachers were able to test simple conjectures within the first few minutes of use with the software. *Fathom*'s ability to easily "drag and drop" variables onto graphs and to be able to link relationships in several graphs made it easy for us to begin using inferential language with teachers from the very beginning (Figure 2).

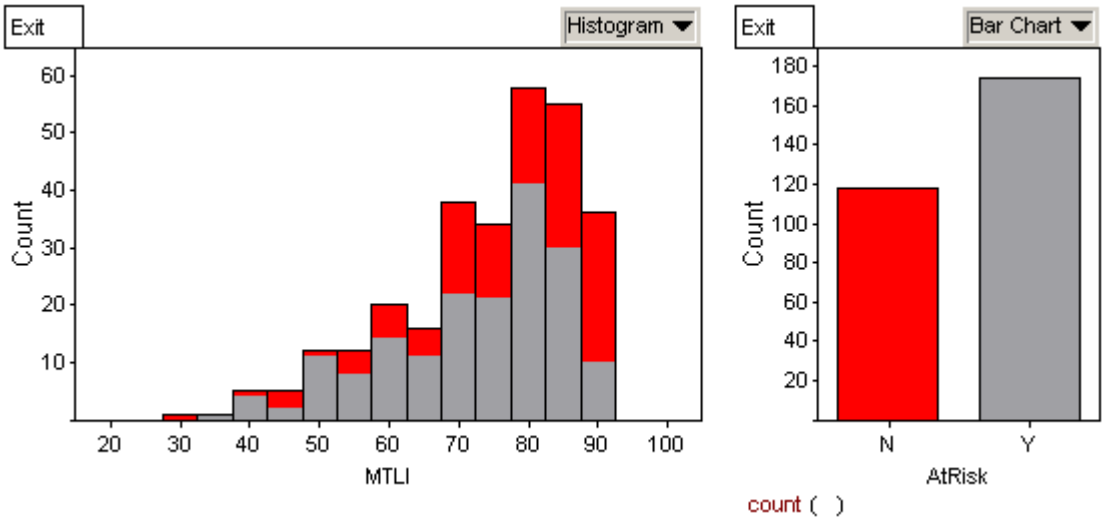


Figure 2: The set of Fathom graphs here are based on data adapted from a local school. The students categorized as "At Risk" (right) have been selected and are highlighted in red within the exit-level math TAAS scores of the all the 10th grade students (left).

We pushed to enlarge teachers' view of data to broaden their perspective, encourage more robust conjectures, expect variation and ambiguity, and look for relationships in the data. As a result, every activity we planned worked towards developing a mindset that would allow for a richer experience with the data. Time was regularly used to discuss issues of accountability, reactions to assigned readings, and general issues of data. As the study progressed, increasing time at the end of the day was dedicated to the teachers' own explorations. The study wove the development of teacher knowledge around four strands: statistical reasoning, investigation into student data, the use of the software, and the process of inquiry.

The findings of the project are given below and detailed in published and presented works (Confrey & Makar, 2001; Confrey, Makar, and Nicholson, 2001; Confrey & Makar, 2002; Makar & Confrey, 2001; Makar & Confrey, 2002; Makar & Confrey, in press):

(1) *Growth in Statistical Content Knowledge.* To measure whether the content that was taught had an impact on teachers' understanding and to assess the level of statistical content knowledge at the time of the interviews, a pre-post test of content knowledge was given to teachers. The result of the analysis is given in Table 1 below. The data summary shows significant growth ($\alpha = 0.05$) in their

overall content knowledge as well as for two individual areas (Sampling distributions and Inference), even though the number of teachers in the study was small (n = 4).

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

Topic	Pre-test Mean Percent Correct	Post-test Mean Percent Correct	Difference	t	p-value
Descriptive Statistics	61%	79%	18%	2.0	0.14
Graphical Representation	75%	83%	8%	0.50	0.63
Sampling Distributions	8%	75%	67%	4.9	<0.01
Inference and Hypothesis Testing	6%	59%	53%	18	<0.01
Overall	35%	71%	36%	6.8	<0.01

(2) *Development of a framework.* A framework was developed (but not tested) that describes five levels of reasoning teachers may use when comparing two groups using four constructs: *creating measurable conjectures, tolerance for variability, integration of context,* and *conclusions/inferences*.

a. At a *Pre-descriptive* level, 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.
b. Teachers using a *Descriptive* level focus on summary statistics and make absolute comparisons between datasets with no regard for variability. Conjectures assume data is infinitely available to answer any question.
c. The first holistic view of the data occurs at the *Emerging Distributional* level, 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.
d. Teachers with a *Transitional View* of the data 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. The concept of statistical *tendency* becomes part of the discussion and conclusion about data.

e. Finally, at the *Emerging Statistical* level, teachers 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 descriptions are well integrated into conclusions and inferences may attempt to draw on statistical models, if relevant.

(3) *Reasoning about comparing distributions*. In examining teachers' reasoning about comparing distributions, we found that teachers were generally comfortable working with and examining traditional descriptive statistical measures as a means of informal comparison. An interesting contrast occurred, however, when we considered teachers' conceptions of *variability* when reasoning about comparing two distributions.

a. The teachers were comfortable in their conceptions of *within-group variation* and were heard in their descriptions of shape, distribution, outliers, standard deviation, range, "domain" (minimum to maximum values), "whiskers" on a boxplot, and statements about a distribution being "tighter" or "more spread out".

b. The teachers typically reported some aspect of the similarity or differences in the measure of *between-group variation*, by comparing the range or standard deviation of each distribution. The teachers often compared shapes or means—for example, by noting that the mean of the females' scores was two points higher than that of the males. Half of the teachers attempted to formally test whether the difference in the means of the two distributions was significant using some form of a standard

deviation taken from the data distributions. A couple of the teachers also checked the size of the population to see if it was 'large enough' to draw samples from. Neither of them, however, used the size of the data set in determining whether the difference in means between the males and females was significant. Overall, the three who considered using a sampling distribution struggled to understand the circumstances under which using one would be helpful nor were they able to separate the variability in the distributions of the data sets from that of the related sampling distribution, confirming that this is a very difficult concept to understand in statistics, consistent with the findings of delMas, Garfield, and Chance (1999).

(4) *Importance of Context for reasoning about variability*. Teachers' accessed different levels of statistical concepts when comparing two distributions with and without a context. Their descriptions of similarities and differences between two distributions with no context were shallow and focused on bar-to-bar comparisons, claims of 'missing' data, and provided little interest in continued discussion. When a context was added to the graphs that was relevant to the teachers (quiz scores from two classes), their descriptions became richer and more statistical. For the first time, they articulated concepts of spread when comparing each pair of distributions and described qualitative differences in the challenges related to teaching classes which were more or less 'together' in a discussion that lasted forty minutes.

(5) *Inquiry*. We found that data-based inquiry was a productive context to work with teachers in addressing the objectives of the professional development. Initially, teachers struggled with moving from a 'topic' of interest to articulating a measurable conjecture. In addition, teachers gained confidence over the course of the workshop with their ability to pursue an independent investigation. They were initially hesitant to be open to the process of being learners, but once they realized that they would not be ridiculed for their lack of statistical knowledge, they became more comfortable with pursuing an investigation and opening up to be learners.

Appendix B: Course Syllabus

EDC 371: Classroom Interactions, Spring 2003
****Special Section on Assessment****
T, TH 9:30-11:00, SZB 316, Unique # 08090

Instructors: Professor Jere Confrey, jere@mail.utexas.edu, 471-1044, SZB 518
Katie Makar, kmakar@mail.utexas.edu, 471-1044 or 232-3958, SZB 518

Office Hours: T & TH 11-12 (or by appointment) in SZB 518 (Southeast side of Sanchez, facing Jester, across from the elevator);

TA: Farhaana Nyamekye

Required Texts:

- Stigler, J. & Hiebert, J. (1999). *The Teaching Gap: Best ideas from the World's Teachers for Improving Education in the Classroom*. New York: Free Press. Available from the Co-op or Amazon.com.
- Heubert, J. & Hauser, R. (1999). *High Stakes: Testing for Tracking, Promotion, and Graduation*. Washington, D.C.: National Academy Press. Available from the Co-op or Amazon.com or read it online at www.nap.edu.

The University of Texas at Austin provides upon request appropriate academic accommodations for qualified students with disabilities. For more information, contact you instructor or the Office of the Dean of Students at 471-6259, 471-4641 TTY.

PREREQUISITES

Classroom Interactions is the fourth course in the UTeach program for secondary math and science teachers. The previous course in the series, **Knowing and Learning, is a prerequisite** for this course. This course builds on experiences from that course. In particular, you should have conducted and analyzed a number of clinical interviews in science and mathematics and be familiar with a constructivist model of learning. If you have not completed Knowing and Learning, you should talk with one of the instructors.

GRADE DETERMINATION

10% Participation: In-class assignments and other participation
20% Take home assignments: lab activities, reflection papers, and other assignments
20% Model Teach: Preparation and implementation of model teaching, including development and analysis of assessment, and reflection paper.
40% Final Project Presentation and Paper: Three-week group inquiry project on assessment and equity using data.
10% Final Exam: Written portion: May 9, 2 – 5 pm. Oral portion, to be scheduled individually.

OVERVIEW OF THE COURSE

The course will be structured into four major sections. All dates below are preliminary and subject to change.

Part 1: Introduction to Assessment. January 14– Feb 6
These three weeks will focus on an introduction to key issues in assessment and accountability. Readings will be assigned from *High Stakes* (NRC, 1999) and *Knowing What Students Know* (NRC, 2001) to introduce you to essential topics in assessment and accountability. In addition, you will be introduced to the software *Fathom* (www.keypress.com/fathom) that we will use throughout the course to examine assessment data.

Part 2: Classroom Instruction (Teaching, Learning, and Assessment). Feb 11 – Mar 6
The next five-week segment of the course will focus on applying ideas of assessment from Part 1 of the course to classroom instruction. During this time, readings from *Teaching Gap* and other readings will be given. In preparation for your Model Teach assignment, you will participate in a number of learning activities in *Fathom* that will both prepare you to teach your lesson and serve as resources for you as you develop your own lesson. Assessment issues discussed in Part 1 will be integrated into this section. This unit will culminate with a 3-day Model Teaching experience in a local high school math or science class (see details below).

Part 3: Test Construction and Analysis. March 18 – April 10
Having completed your Model Teaching, we will come back to take a closer look at classroom assessments and accountability in the context of equity issues. Readings from several sources will be given, as well as additional readings from the *High Stakes* book to examine critical issues in accountability, faced by teachers, their students, and their schools. Data from several sources, including the assessment data you generated from your Model Teaching experience will be analyzed with technological and statistical tools. Further learning activities in Fathom will take you deeper into issues of equity in assessment, providing you with tools to begin to conduct short semi-structured queries into the assessment issues through data analysis. By the end of this segment of the course, you will be fluent in the tools and resources needed to conduct your own inquiry into issues of accountability and equity through assessment data.

Part 4: Inquiry Projects. April 15 – May 1
The capstone of the course will be a project into an issue of equity or accountability that you will investigate and carry out, either alone or with a partner. You will present your findings to the class on one of the final two days of the course. This project will compose a major portion of your course grade and serve to synthesize the readings, teaching and learning experiences, resources, and discussions from the course as well as draw on your experiences during the course and specific interests.

CLASS REQUIREMENTS

Participation & In-class assignments – 10%
The class will typically meet twice per week. Class participation is **required** and will determine a portion of your grade for the course. **Students who are unable to attend class should review Blackboard and contact the TA or the instructor to find out what they missed and negotiate the possibility of making up the work**. There will be a few occasions during the course where your attendance will be required outside of regular class time. Specifically, for your one-day classroom observation, your 3-day model teaching experience, and for scheduled oral interviews. Formative assessments will be given periodically during class to model the importance of embedded assessment.

Take Home Assignments – 20%
Most class sessions will include a follow up and/or preparatory assignment: reading, reflection, lab activity, or other written or oral assignment. **Important**: You must type all written assignments to be turned in, unless you have explicit permission of one of the instructors not to.

Model Teach – 20%
During the course, you will have the opportunity to prepare and teach a 2-3 day lesson on linear regression with a small group of your peers at a local high school. You and your group will be required to develop the lesson plan, execute your plan, and videotape it for your UTeach portfolio. Your group is responsible for checking out the equipment from the LTC, learning how to run the camera, and doing the taping. As a major part of your evaluation of your Model Teach experience, you will be required to prepare, administer, and analyze both a formative and summative assessment. Tentative dates for the Model Teach are for the week of March 3rd.

Final Project & Paper – 40%
The major assignment for this course will be a 2-week inquiry project in April, conducted in pairs. You and your partner will choose from a list of equity and assessment topics and conduct an in-depth, data-based inquiry of a preliminary conjecture you develop. We will develop the skills needed to conduct the inquiry throughout the course. You will present your findings to the class at the end of the course, as well as write a 12-15 page paper articulating the importance of the problem and supporting your findings with evidence.

Final Exam – 10%
You will be given a final exam in this course over the statistical content and inquiry-based reasoning that you learned in the course. Half of this will be a written final exam during the exam period (May 9, 2 – 5pm). The other half will be an oral exam scheduled individually with one of your instructors, of an open-ended problem in data-based evaluation of assessment.

TENTATIVE SCHEDULE

Revisions will be given at the beginning of each Unit. Check *Blackboard* for updates!

Date	In Class	Assignments
	JANUARY 14 – FEBRUARY 6	**UNIT 1: INTRODUCTION TO ASSESSMENT**
Tuesday January 14	Course Overview of syllabus; Research Overview: sign waivers & Pretest	-Begin work on Tutorials 1–3 from the Fathom Workshop Guide available at www.keypress.com/fathom/workshop_guide.html Complete before Jan 23rd.
Thursday January 16	TOPIC: Standards and Assessment ACTIVITY: High Stakes Video, part 1; HANDOUTS: Background on TAAS; Important websites	-Reflection on Video (DO FIRST) -Read *High Stakes* Chapters 2-3 (Chapter 1 optional) -Look at Math, Science, or Technology Standards & TEKS online (bookmark)
Tuesday January 21	TOPIC: Accountability ACTIVITY: High Stakes Video 2 HANDOUTS: TEA Standard-setting; Prado Rulings	-*High Stakes:* Chapter 4 -*Knowing What Students Know:* Chapter 1 -Tutorials 1-3 from *Workshop Guide* due next class
Thursday January 23	TOPIC: Standardized testing ACTIVITY: SAT v. GPA Investigation	-Read *The Great Sorting* -TAAS & ITBS *Fathom* investigation & reflection
Tuesday January 28	TOPIC: Classroom Assessment ACTIVITY: Planning subtraction instruction; Three Video Clips from Confrey teaching subtraction	-*Knowing What Students Know*: Chapter 2-3
Thursday January 30	TOPIC: Formative Assessment ACTIVITY: Examples of Classroom Assessment activities HANDOUTS: NSES & NCTM Assessment Standards; Glossary of assessment terms; Science Educator's Guide to Assessment	-Develop a subtraction lesson & assessment -Reading from NSES or NCTM Assessment Standards
Tuesday February 4	TOPIC: Interpreting Assessment Data ACTIVITY: Examining the Subtraction data HANDOUTS: NSTA Assessment resource; Formative Assessment websites	-Read *Science Educator's Guide to Assessment*; Assessment reflection/observation
Thursday February 6	TOPIC: Linking Qualitative and Quantitative Assessment Analysis ACTIVITY: Structured analysis of Assessment; Examining Student work	-Linking Qualitative and Quantitative Assessment Data
	FEBRUARY 11 – MARCH 6	**UNIT 2: CLASSROOM INSTRUCTION (TEACHING, LEARNING, AND ASSESSMENT)**

Tuesday February 11	TOPIC: Association ACTIVITY: Examining Scatterplots Model Teach Preparation Overview	-*Teaching Gap* Chapters 1-2; -Association Exercises (from Workshop Statistics, Topic 8)
Thursday February 13	TOPIC: Correlation Coefficient ACTIVITY: Examples of Properties and Caveats of Correlation HANDOUT: Frameworks for assessment; Model Teach Expectations	-*Teaching Gap* Chapters 3-4 -Correlation Exercises (WS Topic 9)
Tuesday February 18	TOPIC: LSR in Fathom ACTIVITY: Overview of Modeling & Prediction; Fitting a Line	-*Teaching Gap* Chapters 5-6 -Least Squares Exercises (WS Topic 10)
Thursday February 20	-TOPIC: Residuals -ACTIVITY: Patterns in Residuals	-*Teaching Gap* Chapters 7-8 -Residuals Exercises (WS Topic 11)
Tuesday February 25	-Planning for Model Teach & Assessment	
Thursday February 27	-Dress rehearsal of Model teach	
Tuesday March 4	Model Teaching – NO CLASS	
Thursday March 6	Model Teaching – NO CLASS	
March 8-16	**SPRING BREAK**	
	MARCH 18 – APRIL 10	**UNIT 3: EQUITY AND TEST ANALYSIS**
Tuesday March 18	-TOPIC: Introduction to Equity and Inquiry	-Stats Practice #1 (WS Topic 3 & 4) -Read Case Study of Tree HS (Harvard Paper)
Thursday March 20	-TOPIC: Spread -ACTIVITY: Random Walk; Time to work on Model Teach Paper	[Finish Model Teach Paper]
Tuesday March 25	-TOPIC: Inquiry, Equity, and Data -ACTIVITY: Report back on Model teach;	-Read *High Stakes Testing* one chapter from 5-8 Reflection on either: -Phi Delta Kappan trio articles (Valencia/Scheurich debate) or -TAAS Panel Discussion (March 25, 7pm)
Thursday March 27	-TOPIC: Examining Change -ACTIVITY: Investigating Change	-Read Deficit Thinking chapter and Statesman article on McCallum -Reflection on McCallum Statesman article -Fathom Tutorial 4

Tuesday April 1	-TOPIC: Deficit Thinking -ACTIVITY: *Ignore me* labels	-Read *High Performing Schools* -Stats Practice (WS Topic 5) -Structured Investigation of Tree/McCallum
Thursday April 3	-TOPIC: Taking stock of what we've learned -ACTIVITY: Review of Structured Investigation (using data to be pro-active or re-active)	-Nancy Love article: Glendale, AZ case study -Stats Practice (WS Topic 5) -Semi-structured Investigation: TEA website tour
Tuesday April 8	-TOPIC: Statistics Review -ACTIVITY: Overview of descriptive vs. inferential statistics; Expectations for final projects	-Stats Practice -Preparation for High Stakes Issues Debate – turn in one page of notes pro/con
Thursday April 10	-ACTIVITY: High-Stakes testing debate [Position assigned when you enter class]	-Inquiry project Progress Check -NSES Inquiry Chapters 1-2
	APRIL 15 – MAY 8	**UNIT 4: INQUIRY PROJECTS**
Tuesday April 15	-TOPIC: Begin Inquiry Projects -ACTIVITY: Getting TEA data into Excel & Fathom -HANDOUTS: Possible data Resources for Final Project	
[Wed Apr 16]	Optional Workshop: Comparing Groups (Scrambling in Fathom)	
Thursday April 17	Project Work	
Tuesday April 22	Project Work	
Thursday April 24	Project Work	
Tuesday April 29	Presentations	
Thursday May 1	Presentations	Written Paper Due May 8th 2003 at midnight
FRIDAY MAY 9	FINAL EXAM PERIOD 2 – 5 PM	

Appendix C: Example of Classroom Assessment Analysis using Fathom

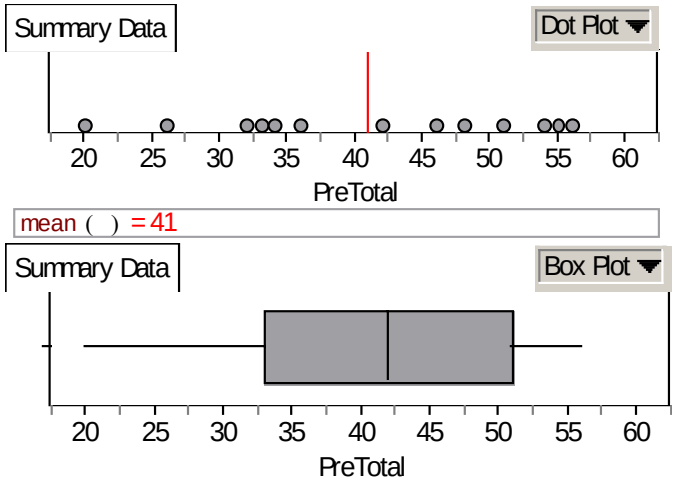
A Structured Analysis of the 6th Grade Subtraction Data

This analysis is organized around the following framework, adapted from the Science Teacher's Guide to Assessment (NSTA, 1998). While not exhaustive, it should provide an example of the kinds of analyses that are possible.

I. Distribution of total student performance.
II. Distribution of student performance on related subsets of items
III. Distribution of difficulty of questions
IV. Examining individual student work
V. Slicing the Data

I. Distribution of total student performance

PRETEST



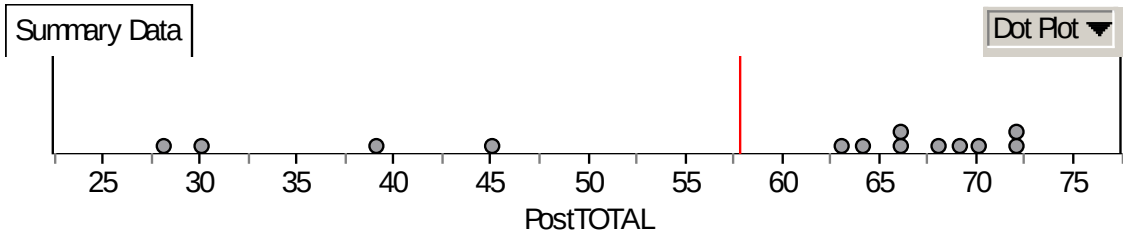
Overall performance on the pretest was extremely variable and fairly evenly distributed, as seen in the dotplot; students answered from 20 to 56 questions correctly out of a possible 59 with an average score of 41-42. The mean is a reasonable summary of the overall performance of the class, but the spread of the distribution is a critical indicator to the teacher that they will have a challenge in keeping all students engaged in the unit. The boxplot on the right gives a quick overall perception that there is little clustering of scores on the whole as the lower three quartiles are each fairly wide relative to the range of scores, and the middle 50% of scores are fairly symmetric.

There is a clear need for differential treatment of students on this unit. Given the size of the class, it is reasonable to investigate individual student work for a significant proportion of the class. Depending on the resources, instruction could be targeted at students in the middle (scoring 32 – 48 points), with additional work for the upper and lower groups (three or four at the top and two at the bottom). Three of the students (Alan, Ana, and Brice) answered nearly all of the questions correctly and need more challenging material from the beginning. It's possible that the student scoring 51 (Nala) would also be included in this group – I would want to look at her work more closely to determine which questions she missed and whether any systematic or consistent errors exist in her work. Given that she is a high performing student, I anticipate her weaknesses to be minor and would probably include her with the other three. She would probably get the additional practice she needed through more advanced work.

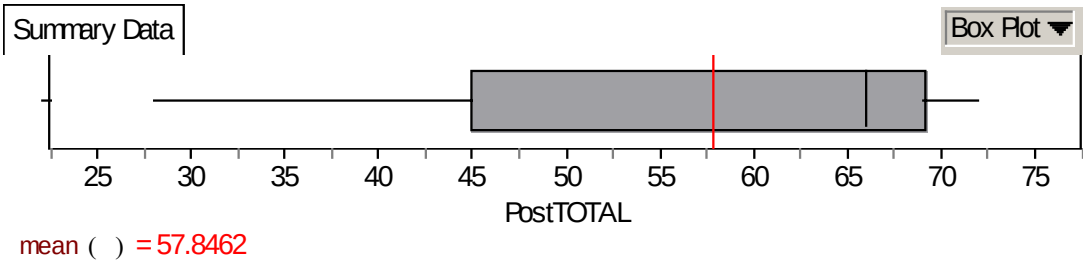
On the other end, two students (Mari & Ellen) need extensive work on their understanding of addition and subtraction of integers, and their performance is clearly lower than the others in the class. I would examine their work more closely to determine whether the low scores are due to systematic errors (e.g. error due to sign) or whether they left a number of questions blank (perhaps due to lack of confidence).

There is a large gap in the middle of the class and indicates another possible division. At this point, the teacher would need to consider his/her goals for the unit and whether these goals are best accomplished with students working on mostly the same material or whether students need to be separated.

POSTTEST



The dotplot of the posttest shows a clear picture – all but 4 of the students (Mari, Ellen, Celeste, and Madden) performed quite well, scoring between 63 and 72 questions correctly out of a possible 73. The mean, 58 questions correct out of 73, is not a reasonable indicator of overall performance given the distance of the lowest four scores from the rest of the class – these scores are skewing the mean downwards. The boxplot (below) is somewhat misleading as the small size of the class makes the third quartile appear very spread out, while most of the space in the third quartile (in the range of 45 to 66) is empty. It does indicate, however, a high median score of 66 questions answered correctly, more indicative of overall performance of the class. The discrepancy between the mean and median further indicate how skewed the distribution is to the left and signals the existence of a few extreme low scores.

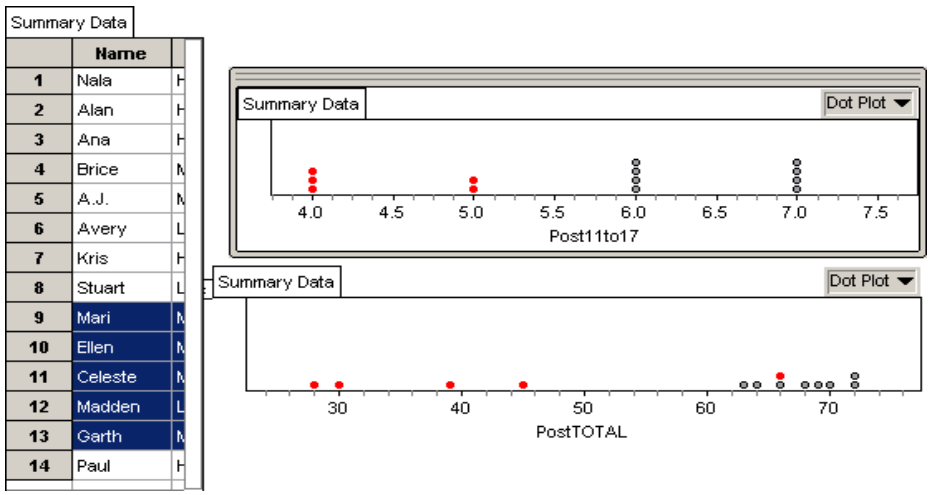


Instructionally, the four students who performed very poorly (answering about 40-60% of the questions correctly) will need additional support, but as a class students appear ready to build on the understanding they have developed.

II. Distribution of performance on related subsets of items

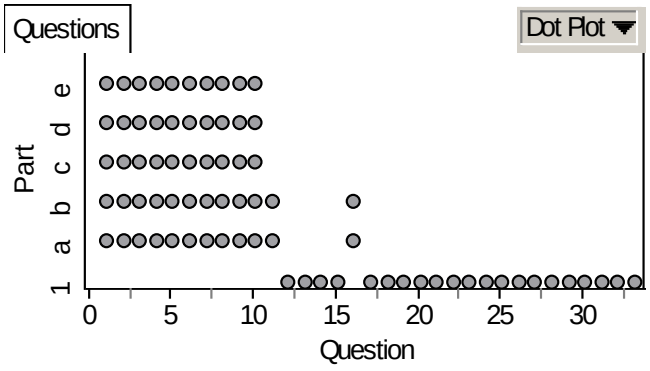
One possible analysis to be done here is to examine the performance of students on the applications portion of the posttest as a follow-up to the whole test analysis above. Because the skills portion makes up such a significant number of the questions, the four students who performed poorly would have had to struggle on the skills portion, while those who did well on the test most likely were successful on this portion. However, I am left with greater uncertainty about the performance on the middle portion of the test – questions 11-17. Given that #14 is a skills question, I may choose to remove this item from the applications analysis. However, because of the way the data is organized, I can get close enough general information by leaving it in and is probably not worth the additional effort to remove it at this level of analysis.

The middle section, questions 11-17 consists of 8 questions. There was no #13 on the posttest, but questions 11 and 16 had two parts each.



The dotplots above indicate an interesting outcome for the applications problems. Although there were 8 questions, no one got all 8 correctly. I would be curious as to whether there was one particularly problematic question that nearly no one answered correctly, or whether the errors were fairly evenly distributed across the eight questions. [Further analysis indicates that no one answered question 17 correctly, where students had to create a scenario for 5 – (-7).] I am surprised, given the clustering of scores on the posttest, that the performance on these items is so uniform, with all students answering between 4 and 7 out of 8 questions correctly. One would expect, given the gap in performance between the bulk of the class and the four who struggled, that these scores would also have been more variable. By selecting the students who scored below average on the applications problems, we can see in the set of graphs and table below that our four struggling students are in this lower half, as might be anticipated, but also that one of the lowest performers, earning 4 out of 8 applications questions, was also one of the higher scoring students (Garth) on the overall test. I would be concerned that this student's high performance was due to significantly strong skill-level work and once the questions were weighted, this student would not have performed as well on the test overall. The four low-scorers on the overall test, however, did no worse on the applications section, usually considered more difficult, than the skills section of the test, a finding that surprised me but that indicates to the teacher that students who struggle with skills could be no less capable on applications. Often teachers assume that if students are unsuccessful at skills, they will not be able to do application problems. As a note, it will be interesting to later investigate the performance of skills v. applications questions. I also wonder whether the students who did poorly answered the same 4-5 questions correctly or whether their performance on the applications section of the test is more scattered.

III. Distribution of difficulty of the questions

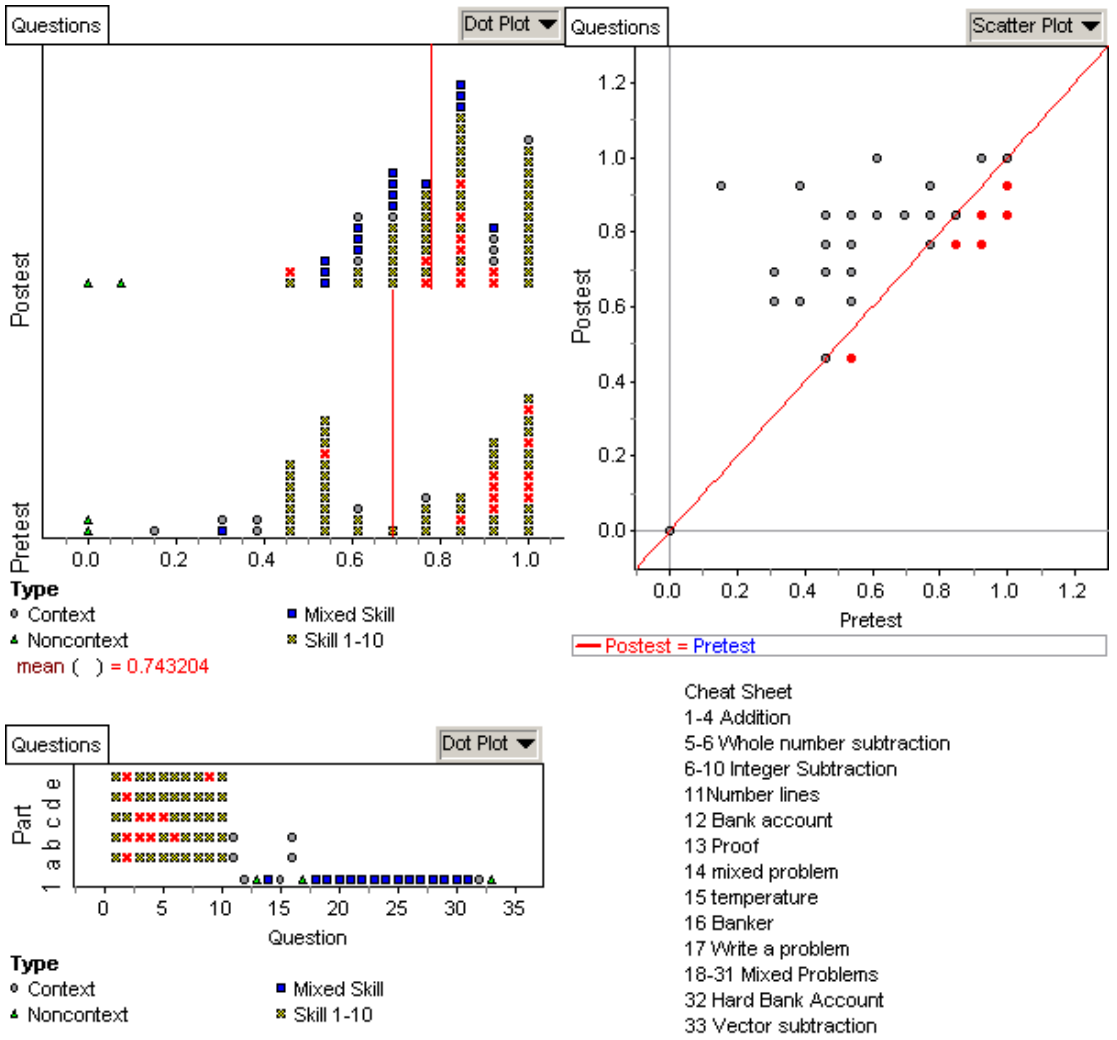


When setting up the data set, I created fields that would allow me to easily identify separate parts of the test. For example, in the dotplot below, I was able to separate out each question so that when I looked for relationships in the difficulties of the questions, I would not have to go to the table to figure out which questions selected segments of

the distribution applied to. Also, at this point, as I get into more detailed analysis, it is difficult not to dig further into what I am seeing. My analyses are much more relational in nature, rather than simply descriptive at this level, as can be seen by the representation set on the next page.

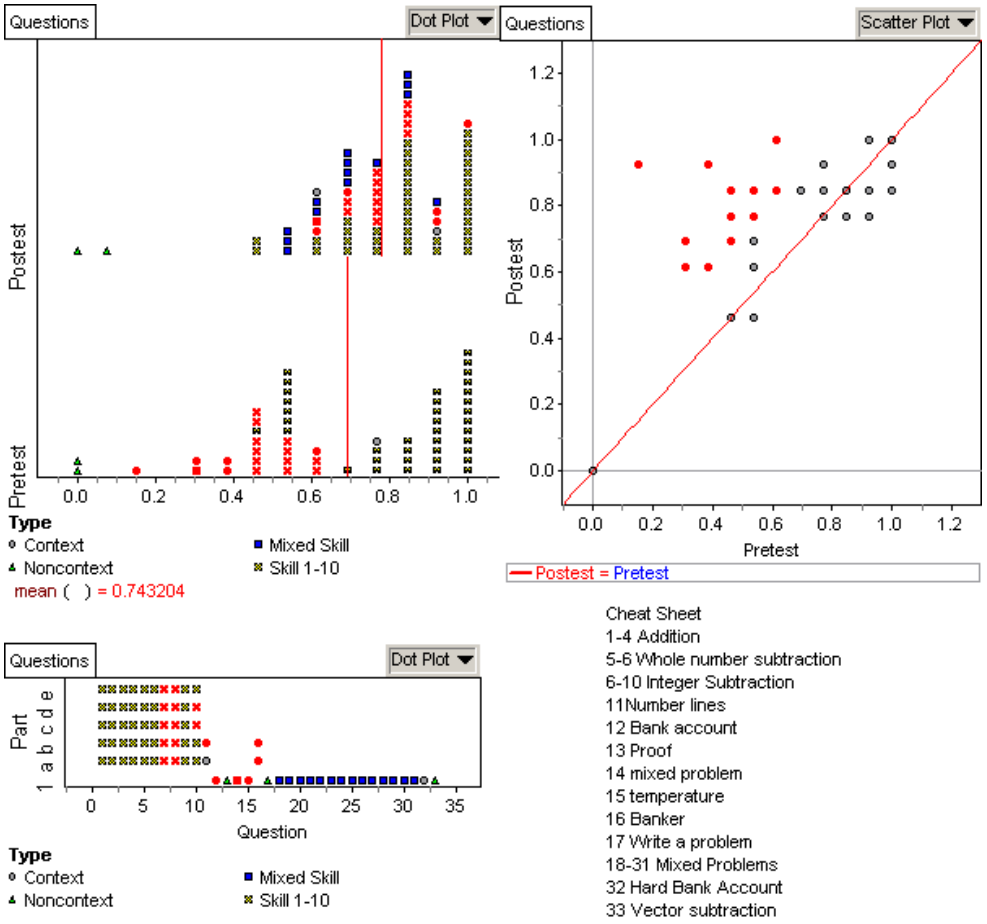
The figure below (top of page 5) has four parts. The top left graph displays the pretest and posttest questions on a stacked dotplot, each dot on the graph represents one question on the test. The scale, ranging from 0 to 1, indicates the percentage of students who answered each question correctly. I added an attribute to the data set "Type" with four categories indicating whether each question was a routine skills-based question "Skills 1-10" (from questions 1-10), a contextual (application) question "context" (from questions 11,12,15,16, or 32), or a noncontextual question "noncontext" (#13,17,33). This graph shows all of the questions, not just the common ones. This new attribute "Type" was dropped into the middle of the stacked dotplot so that I could quickly determine the types of questions that were appearing in my analysis.

The bottom left graph is the graph shown at the beginning of this section – it simply provides a quick graphic organizer of the questions on the test so that I can look for patterns among the questions. (I also added a "cheat sheet" text box on the right of this graph to remind me of the content of the questions). As above, I dropped the "Type" attribute into the middle of the graph to see which kind of question might be selected (skills 1-10, mixed skills, context, or noncontext).



The top right graph is a scatterplot with the difficulty of the pretest and posttest questions (common). The line Pretest = Posttest indicates the division where questions were the same level of difficult on the Pretest as the Posttest. Questions below this line indicate a *drop* in performance; that is, these questions were missed more often (students did worse) on the Posttest than on the Pretest. These questions are clearly a concern and have been highlighted. Note that all of the questions where student performance decreased come from the skills section, questions 1 – 10, as seen in the graph on the bottom left. In fact, all but one (Question 9e) are from questions 1 – 6, which are either addition questions or subtraction of positive integers. With one exception (again, Question 9e), the performance on these questions were all quite high on both the Pretest and Posttest, indicating that the drop was probably just due to careless errors, although that might be followed up by a more qualitative analysis. Note that all of the questions that showed a drop in performance showed only a minor drop, indicated by the fact that while these points lie below the line, they do not lie far from the line. That is, the performance on the Pretest and Posttest were fairly close.

Questions above the Pretest = Posttest line (see below) indicate the questions on which student performance improved – the farther the point is above the line, the greater the improvement. If we look at questions where students improved the most, we see that none of these questions came from #1-6. Interestingly, students improved significantly on all parts of questions 7 & 8. In addition, almost all of the contextual problems are in this group, indicating that instruction on application problems was the most successful.

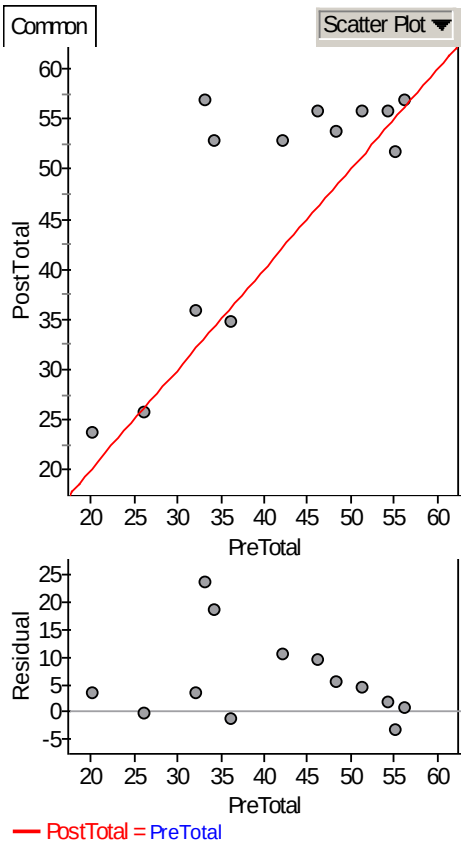


IV. Individual Student Work

A qualitative discussion, by examining student work, is given in class.

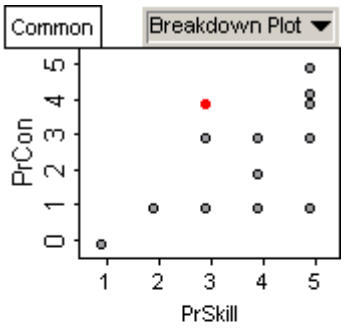
V. Slicing the Data

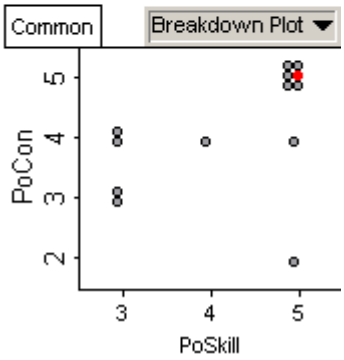
This is where data analysis allows us to look at slices of the data that are unique and insightful. In this section we will examine three different "slices" of the subtraction data set to mine for relationships. First, we will look at the performance of items that are contextually-based and their related non-contextual versions. Next we'll examine the performance of students of students on subsections of the test relative to their overall performance. Finally, we'll look at relative difficulty of the skills problems in questions 1 – 10 to see if there is any difference in performance relative to the level of difficulty of the questions.



In the graph at left, we see a scatterplot of students' pretest total score along the horizontal axis (PreTotal) and their score on the posttest of items common on both tests along the vertical axis (PostTotal). The line PostTotal = PreTotal (y = x) is plotted to demarcate students who improved on the posttest (above the line) and those who did not (on or below the line). We already examined the … Of particular interest might be two groups of students: (1) the four who did poorly on both the pretest and posttest, and (2) the four students who showed the greatest overall improvement on the test (largest residuals).

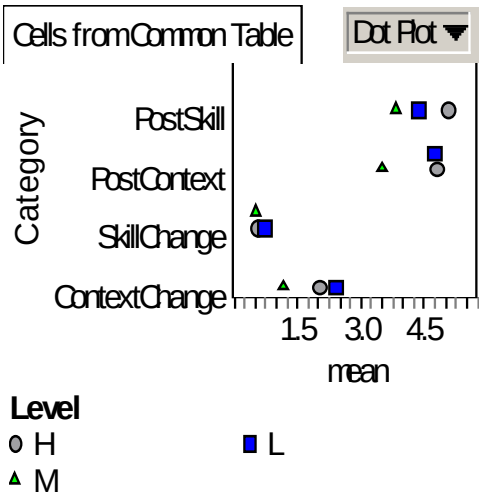
Begun above, additional analyses could look at comparing performance on contextual problems vs. skills problems (especially the link between the two pairs of problems where the same numbers were used in a context and non-context situation), routine vs. mixed skills problems and a closer examination of performance on specific questions.





All but three students performed better on the skills set than the context set on the pretest, but on the posttest, only two students did, indicating a greater overall growth on the contextual version of the problems than on the skills version. Furthermore, if we examine the performance of the lowest third of the class, we find that these students improved more, in general, than the rest of the class (greater mean change), indicating that the contextual problems not only moved the class up in performance, but also worked towards closing the gap between low achieving students and the rest of the class.

The representation at right shows the mean number of questions answered correctly on the posttest for the skills set and the context set. In addition, we see the mean change in questions answered correctly from the pretest to the posttest on the 5 pairs of questions for which the same questions were asked in a skills-only context and its corresponding question in a contextual setting. The representation also shows that if we examine the performance of the students identified as low achievers in the course, they not only performed well on the posttest relative to the middle and high achievers, they lead the class in improvement in both settings with the greatest improvement in the contextual problems.



Appendix D: Posttest

Classroom Interactions Posttest

Survey Information NAME:_____

The purpose of this survey is to find out the impact that this course may have had on your thinking. It is not a course evaluation (you've already done that), but an opportunity for me to find out in more detail what you have learned in this course. Please answer as honestly as you can.

A. Statistics Background. Please describe your statistics coursework/experiences you've had outside of this class. Include major topics studied.

B. Overall Course. When you think back on this course <u>overall</u>, what stands out as the most meaningful experience(s)/assignment(s) for you personally. Please explain and provide an example.

C. **Course Themes.** This course has had four major themes: Equity, Assessment, Data/Statistics, and Inquiry. For each theme, please describe what you think has been the most useful/meaningful idea (intellectually or personally) that you have learned in that theme. Include ideas that you think you will carry with you beyond the course. If you need more space, please use extra paper.

1. Equity

2. Assessment

3. Working with data/learning statistical concepts

4. Inquiry

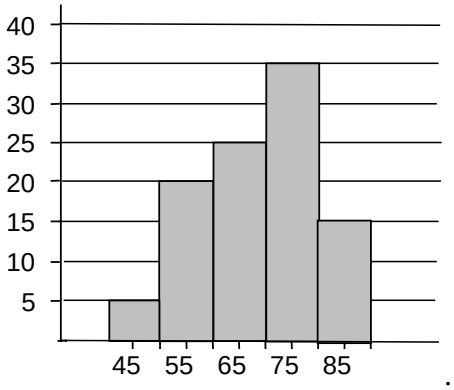
D. Using Data to learn Equity. Has working with authentic data and/or learning statistics during this course helped you to better understand equity? If so, how? If not, did it detract or just not affect your understanding? In either case, please explain.

E. **Statistics Comfort level**. Please rate your level of comfort with each topic below with **1** ***being very low/none*** and **5** ***being 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 Thm, etc.)	**1 2 3 4 5**
Statistical Inference (t-tests, confidence intervals, chi-square tests, power, Type II error, ANOVA)	**1 2 3 4 5**

II. Statistics and Data Questions. Please provide the best answer for each of the following questions. Note that starred items (*) are extra credit.

Questions 1 – 7 refer the histogram below showing the test scores for a group of students.



1. What do the numbers on the horizontal axis represent?

   (A) The independent variable
   (B) Test scores
   (C) Number of students with a test score in each interval
   (D) The dependent variable

2. What do the numbers on the vertical axis represent?

   (A) The independent variable
   (B) Test scores
   (C) Number of students with a test score in each interval
   (D) The dependent variable

3. How many students have test scores below 60?

   A) 1 B) 2 C) 20 D) 25 E) 35

4. How many students are represented in the graph?

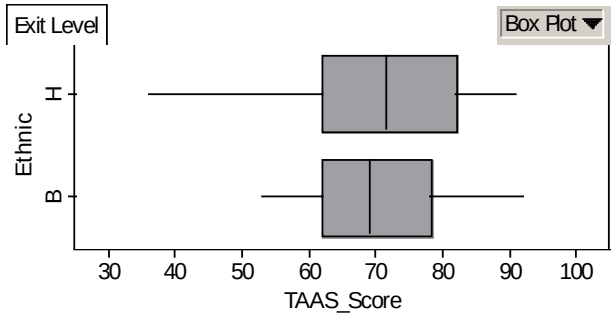
   A) 5 B) 35 C) 50 D) 90 E) 100

5. Calculate the proportion of students measured with test scores below 60. Show work where needed.

6. Use the graph to estimate the median test score for this group of students. Show work where needed.

7. Use the graph to estimate the mean test score for this group of students. Show work where needed.

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.



Group	Number of Students	Mean TAAS score	Percent Passing TAAS
African-American	31	71.0	48.4
Hispanic	228	71.5	61.4

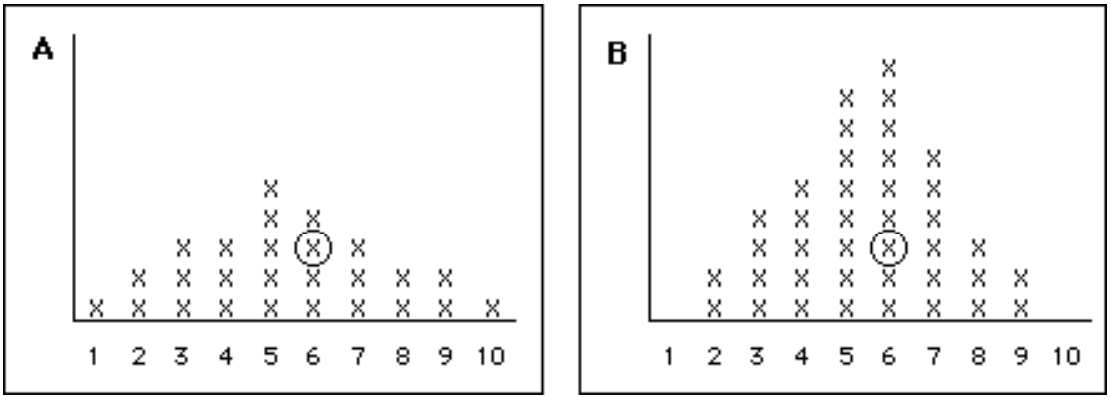
8. List at least three conclusions that would complete the following sentence: "By comparing the performance of Hispanic students with the performance of African-American students, I would draw the following conclusions…"

9. Given the information above, list two recommendations would you make to your principal to help ensure that the school is does not get identified as low performing again in the following year.

10. If the principal takes your advice, what data would you want to collect during the next year in order to monitor the success of your proposals.

11. A six-sided die is thrown 7 times resulting in the following outcome: 3, 3, 3, 4, 4, 5, 5 (order is not important). Do you think there is evidence to suspect that the die is unfair? Why or why not?

12. In the graphs below, Figure A represents a distribution of 26 weights (rounded to the nearest kilogram). Figure B represents a sampling distribution of mean weights (rounded to the nearest kilogram) for samples of size 3. One value is circled in each distribution.

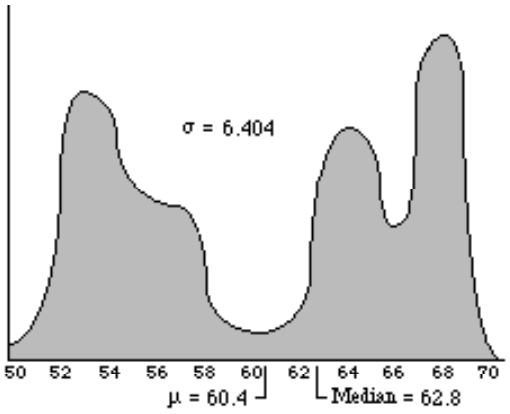


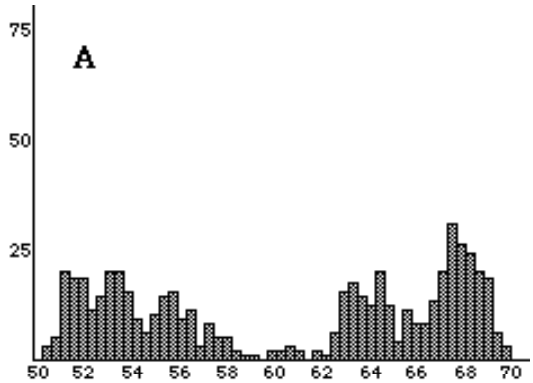
Which statement below best describes the comparison between what is represented by the X circled in Figure A and the X circled in Figure B?

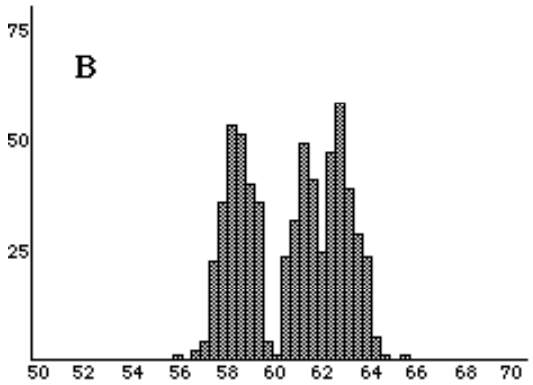
(A) They both represent 6 kilograms, so there is no difference.
(B) There is a difference, because there are more values in Figure B than in Figure A.
(C) There is a difference. In Figure A, the X represents a single weight, but in Figure B the X represents the mean of three weights.
(D) There is a difference because the graph in Figure B is more like a normal distribution than the graph in Figure A.
(E) There is a difference because Figure A is centered around a value of 5 but Figure B is centered around a value of 6.

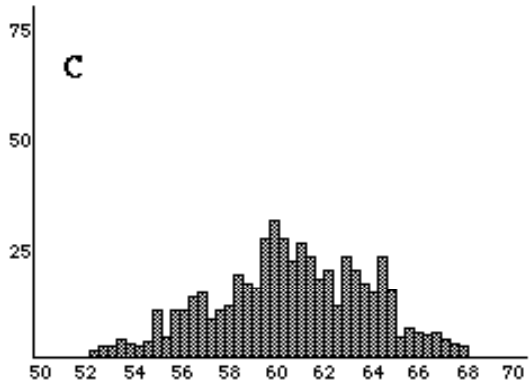
Questions 13 – 19 on the next page refer to the graphs below. The distribution of a population of test scores is the first graph, labeled POPULATION**,** displayed below. Each of the other five graphs labeled A to E represents a possible distribution of sample means for random samples drawn from the population. Answer the questions on the next page about these figures.

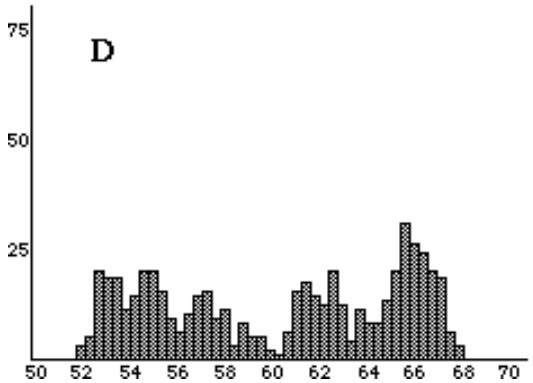
POPULATION

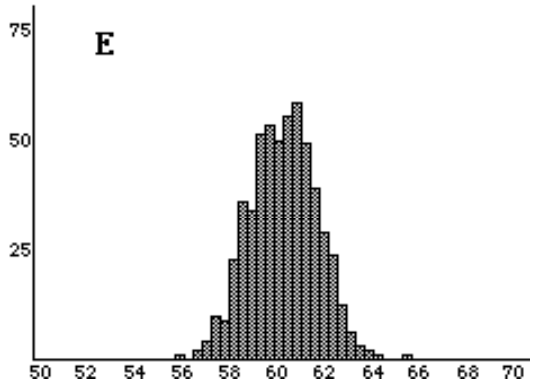












In Questions 13 – 15, 500 samples of **size 25** are randomly drawn from the POPULATION distribution.

13*. Circle the letter that best represents a graph of this distribution of sample means.

A B C D E

14*. I would expect the sampling distribution to be shaped more like:

A) a normal distribution
B) the population

15. Which phrase comes closest to 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.

In Questions 16 – 18, 500 samples of **size 4** are randomly drawn from the population.

16*. Circle the letter that best represents a graph of this distribution of sample means.

A B C D E

17*. I would expect the sampling distribution to be shaped more like:

A) a normal distribution
B) the population

18. Which phrase comes closest to 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.

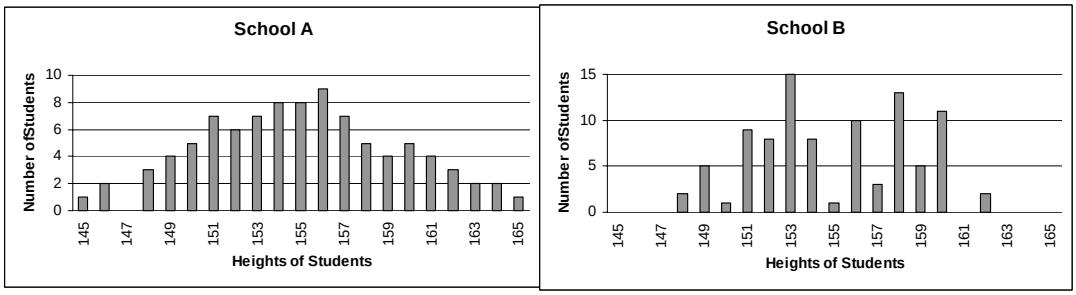
19. Which phrase comes closest to completing the following sentence?
I expect the sampling distribution in Question 13 to have…

(A) less variability than the sampling distribution in Question 16.
(B) the same variability as the sampling distribution in Question 16.
(C) more variability than the sampling distribution in Question 16.

20*. Weight is a measure that tends to be normally distributed. Suppose the mean weight of all women at a large university is 135 pounds, with a standard deviation of 12 pounds. In a random sample of 9 women at the university, there is a 68% chance that the sample mean weight would be between:

A) 119 and 151 pounds.
B) 125 and 145 pounds.
C) 123 and 147 pounds.
D) 131 and 139 pounds.
E) 133 and 137 pounds.

21.



The graphs above describe some data collected about Grade 7 students' heights in two different schools. Which graph shows more variability in students' heights? Explain why you think this.

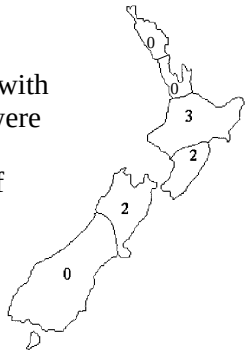
22. Given the average summer temperature in cities P and Q, explain briefly how you would decide which of the following two events is more unusual: a 90 degree summer day in city P or a 90 degree summer day in city Q.

23. 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)

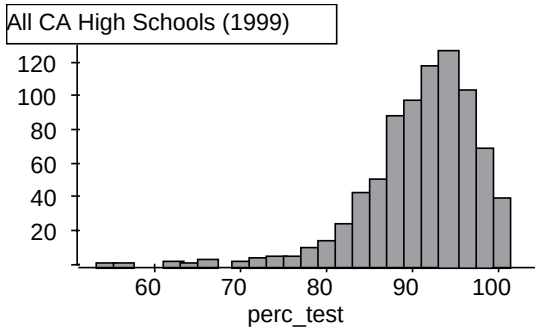
24. When a set of data has suspect outliers, which of the following are preferred measures of central tendency and of variability?
    A) mean and standard deviation
    B) mean and variance
    C) mean and range
    D) median and range
    E) median and interquartile range

25. Every year in New Zealand approximately seven children are born with a limb missing. Last year the children born with this abnormality were located in New Zealand as shown on the map below. Note that the population in each region below is approximately equal. A group of families in the central regions have filed a legal case claiming the incidence in their region is unusually high. Do the data support their claim? Why or why not?



26. The graph below right gives 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

27. 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.



Appendix E: Inquiry Project Assignment

Classroom Interactions (Special Section)
Final Project Presentation and Paper

Presentations: April 29 & May 1 (in class)
Paper Due: May 8 (midnight)

The capstone of the course is a 2-week inquiry project examining an issue of equity or accountability that you will investigate and carry out – you may work alone or with a partner. You will choose an investigation important to you around a question of equity, and conduct an in-depth, data-based inquiry of a preliminary conjecture you develop. You will present your findings to the class on one of the final two days of the course (April 29 or May 1), as well as write a 12-15 page paper (not including appendices) articulating the importance of the problem and supporting your findings with evidence. This project will compose 40% of your course grade and serve to synthesize the readings, teaching and learning experiences, resources, and discussions from the course as well as draw on your experiences during the course and specific interests.

Your inquiry should be structured after one of the following types of investigations, or you may propose an alternative (which must be approved in advance). Depending on the conjecture under inquiry, you may choose to combine more than one type of investigation if it provides additional insight into your topic of study.

Types of Investigations

- Description study – using data to support a specific issue under investigation
- Comparison study – across race, ethnicity, gender, socioeconomic status, performance level, language, disability, or item type
- Longitudinal analysis of data
- Correlation study comparing associated variables
- Analysis of test items in relation to construct validity

Your paper should contain six main sections:

1. Introduction (1-2 pages) – Statement of the problem, introducing the area you choose to investigate and why it is of interest to you.
2. Link to Equity (2-4 pages) – Further discussion of the problem you are investigating and why it is important, using the readings from the course to support your investigation. Additional readings to support your research are encouraged – see the instructors for suggestions tailored to your area of interest.
3. Method of study (1-3 pages) – Process used to conduct your investigation including a description of initial and refined conjectures and how you went about your investigation.

4. Data and Results (3-5 pages) – Description (with graphical displays) of the data found and the results of your investigation. Your descriptions here should be unbiased – that is, you should describe your statistical analysis in relation to the context, but do not discuss the implications of your findings until Section 5.
5. Discussion and Conclusion (4-6 pages) – Interpretation of your results and their implications. This should is where you should make any arguments for potential meaning of your results, tying your findings to your equity discussion in Section 2. Potential shortcomings of your study should be mentioned here as well as ideas for further investigation based on your findings. Be sure to end with a short conclusion summarizing your findings and a list of references.
6. Appendices – Include here any data or graphs that were too large to include in the text above, Fathom files (organized), or websites used to find your data. The pages here should not be disconnected from your paper. That is, they should only be included if referred to in the text of your document and/or annotated and have a clear purpose. Avoid attaching pages of documents or files that are not of use to the reader.

Your grade will be based on the level of your analysis, organization and discussion of your investigation, use of evidence (data and readings from the course) to support your findings, and ability to tie together your experiences and resources drawn from the course.

You will also provide a 15-minute presentation to the class on April 29 or May 1 that highlights the findings of your analysis and then lead a 5-minute question and discussion session. Let your instructors know in advance if you will need computer and projection equipment for your presentation.

Examples of Studies

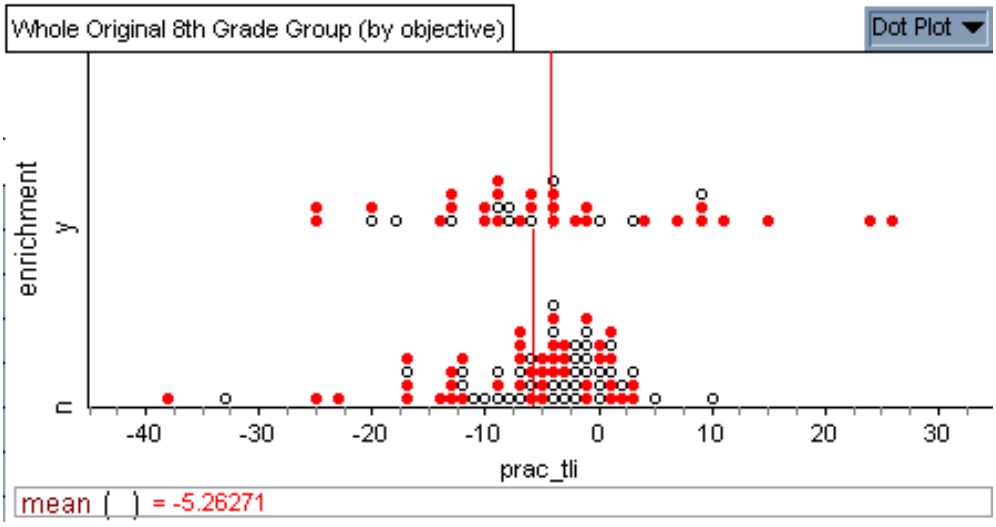
- A specific equity issue that you would like to explore using data from TAAS, End-of-Course exams (Biology or Algebra), SAT, NAEP, school finance reports, or other data (for example, on the TEA website), with a particular focus on one group (e.g. gender, LEP, ethnicity) or a specific inquiry regarding the test.
- An analysis of the items and related data on TAAS or an end-of-course exam and how the data can be used to give teachers feedback.
- An investigation of equity issues through psychometric analysis of TAAS (you will need to learn an analysis program called *Winsteps*), possibly working alongside graduate students. For example, for students with the same test score, is there a difference in the types of questions females find easier than males (and vice versa)?
- A look at potential instructional treatments in the ways that these might impact student performance over time, using data to support your analysis
- A case study of a particular school or district, using data to support your analysis
- A study of high stakes testing across several states

An annotated list of potential web resources will be provided.

Appendix F: Post-interview Questions

Part I. Offline Questions (Used for pre- and post-interviews)

Background information:
A local urban middle school has created a program to assist students who need extra help to prepare for the TAAS Math exam. They meet as a regularly scheduled class called "Math Enrichment". The students were placed in the class if their counselor determined they needed it. The school is interested in whether the program is helping students to improve their scores and have collected data on the difference between their 7th grade TAAS math test and a practice TAAS test given to them in the spring of the 8th grade. A graph of the change in scores is shown below [Show them enlarged graph]. A positive difference indicates that the student scored BETTER on the 8th grade practice math test than on the 7th grade math TAAS. The data from the students in the enrichment class are on the top distribution and the data from the students not in the enrichment class are on the bottom distribution. The mean improvement of each group is shown on the graph. In addition, students highlighted are classified as Economically Disadvantaged.



1. Compare the improvement of the students who were in the enrichment class with those who were not.

2. In your opinion, is the program working? Should they keep it? Evidence??

3. Do you have any concerns about the equity of the program?

4. Here is a list (show list, on separate page) of additional data that was collected about the students in the 8th grade. [Go through each attribute and describe it].
   a. Using this data, or other data you might collect, what is a question that you might ask to explore the issue further of whether the program is working and/or equitable?
   b. Can you state a conjecture or hypothesis you might have about the question you are asking?

EconDisadvantaged	If student has applied for free or reduced lunch
enrichment	If they've been placed in a TAAS remediation class
ethnicity	The ethnicity or race of the student
SexCode	Gender
MAPYTLI	6th grade Math TAAS score (TLI)
MATLI	7th grade Math TAAS score (TLI)
P1MTLI	fall 8th grade score on a district Math TAAS practice test
P2MTLI	spring 8th grade score on a district Math TAAS practice test
Prac-TLI	Improvement in their score from their 7th grade Math TAAS to their spring 8th grade math practice TAAS.
REPYTLI	6th grade Reading TAAS score
RETLI	7th grade Reading TAAS score
PracRE	8th grade score on the district reading TAAS practice test

Part II: Online investigation (Used for post-interview only)

[Note: Turn on the computer recorder]

On the computer desktop is the Fathom file *Hispanic Urban and Rural.ftm* of a random sample from Texas of Hispanic students who live in Urban or Rural areas. [check that they know what Urban and Rural mean].

1. Before you open the file, make a conjecture about the performance of Hispanic students in Urban vs. Rural areas.
2. Open the file and investigate your conjecture.
3. What did you find? What would you conclude?
4. Based on what evidence?
5. Why do you think the rural students did better (or the same)?
6. Check Economically Disadvantaged status (breakdown plot) – probe reasoning
7. Do you see any equity issues that need to be addressed for Hispanic students in urban vs. rural schools?
8. Now compare the two distributions *statistically* without thinking about the data coming from test scores (prompt: suggest using the ways of describing distributions that we used in class)

References

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

Allexsaht-Snider, M., & Hart, L. (2001). "Mathematics for all": How do we get there? *Theory into practice, 40*(2), 93-101.

American Association for the Advancement of Science. (1989). *Project 2061*. New York: Oxford University Press.

Apple, M. W. (2001). *Educating the "right" way: markets, standards, God and inequality*. New York: RoutledgeFalmer.

Bakker, A. (2001). From data via 'bump' to distribution. Presented at the Second International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL2); Armidale, Australia.

Bakker, A. (2002). *Route-type and landscape-type software for learning statistical data analysis.* Paper presented at the Sixth International Conference on Teaching Statistics (ICOTS6), Cape Town, South Africa.

Bakker, A. (2004). *Design research in statistics education: On symbolizing and computer tools.* Freudenthal Institute, Utrecht.

Bakker, A. (in press). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Ball, D. L. (1996). Connecting to mathematics as part of learning to teach. In D. Shifter (Ed.), *What's happening in math class? Reconstructing professional identities* (Vol. 2, pp. 36-45). New York: Teachers College Press.

Ball, D. L. (2002). *Mathematical proficiency for all Students: Toward a strategic research and development program in mathematics education*. Santa Monica, CA: RAND Education/Science and Technology Policy Institute.

Batanero, C., Garfield, J., Ottiavani, M. G., & Truran, J. (2000). Research in statistical education: Some priority questions. *Statistical Education Research Newsletter, 1*(2), 2-6.

Batanero, C., & Serrano, L. (1999). The meaning of randomness for secondary school students. *Journal for Research in Mathematics Education, 30*(5), 558-567.

Becker, H. J. (2000). Internet use by teachers, *Technology and learning* (pp. 80-111). San Francisco: Jossey-Bass.

Begg, A., & Edwards, R. (1999). *Teachers' ideas about teaching statistics.* Paper presented at the joint conference of the AARE & NZARE, Melbourne.

Ben-Zvi, D. (2000). Toward understanding the role of technological tools in statistical learning. *Mathematical Thinking and Learning, 2*(1&2), 127-155.

Ben-Zvi, D., & Arcavi, A. (2001). Junior high school students' construction of global views of data and data representations. *Educational Studies in Mathematics, 45*, 35-65.

Bernal, E. (2000). Psychometric inadequacies of the TAAS. *Hispanic Journal of Behavioral Sciences, 22*(4), 481-507.

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

Biehler, R. (2001). Developing and assessing students' reasoning in comparing statistical distributions in computer-supported statistics courses. Presented at the Second International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL2); Armidale, Australia.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

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

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school (expanded edition)*. Washington, D.C.: National Academy Press, National Research Council.

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.

Brush, S. (1991). Women in science and engineering. *American Scientist, 79*, 404-419.

Burrill, G. (1996). *Discussion: How technology is changing the teaching and learning of statistics in secondary schools.* Paper presented at the Research on the Role of Technology in Teaching and Learning Statistics, Granada, Spain.

Burrill, G. (1997a). Graphing calculators and their potential for teaching and learning statistics. In J. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics*. Voorburg, The Netherlands: International Statistics Institute.

Burrill, G. (1997b). Personal communication. Kuala Lumpur, Malaysia.

California Commission on Teacher Credentialing. (2000). New technology standards for teachers, *Technology and learning* (pp. 33-37). San Francisco: Jossey-Bass.

Campbell, P. F., & Silver, E. A. (1999). *Teaching and learning mathematics in poor communities*. Reston, Va.: Task force on Mathematics Teaching and Learning in Poor Communities: National Council of Teachers of Mathematics.

Canada, D. (2004). *Preservice teachers' understanding of variation.* Unpublished doctoral dissertation, Portland State University, Portland.

Casey, B. M. (1996). Understanding individual differences in spatial ability within females: A nature/nurture interactionist framework. *Developmental Review, 16*, 241-260.

Castro-Filho, J. A. (2000). *Teachers, math, and reform: An investigation of learning in practice*. Unpublished doctoral dissertation, University of Texas at Austin.

Catsambis, S. (1994). The path to math: Gender and racial-ethnic differences in mathematics participation from middle school to high school. *Sociology of Education, 67*(3), 199-215.

Chance, B., Garfield, J., & delMas, R. (2001). Developing simulation activities to improve students' statistical Reasoning. Prereading for the Second International Research Forum on Statistical Reasoning, Thinking, and Literacy, Armidale, Australia.

Clewell, B. C., & Campbell, P. B. (2002). Taking stock: Where we've been, where we are, where we're going. *Journal of Women and Minorities in Science and Engineering, 8*, 255-284.

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

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9-13.

Cohen, D. K., & Hill, H. C. (2001). *Learning policy: When state education reform works*. New Haven: Yale University Press.

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

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

Confrey, J. (2000). Function probe. Austin, TX: Quest.

Confrey, J. (2002a). *Design research*. Unpublished manuscript, Austin, TX.

Confrey, J. (2002b). Open letter to state board members. Austin, Texas. www.syrce.org.

Confrey, J. (2003). Systemic crossfire: The students speak out (Video). Austin, Texas: Systemic Research Collaborative for Education in Mathematics, Science, and Technology.

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

Confrey, J., Bell, K., & Carrejo, D. (2001). *Systemic crossfire: What implementation research reveals about urban reform in mathematics*. University of Texas at Austin, www.syrce.org.

Confrey, J., & Carrejo, D. (2002a). *A content analysis of exit level mathematics on the Texas Assessment of Academic Skills: Addressing the issue of instructional decision-making in Texas.* Paper presented at the Twenty-fourth Annual Meeting of the North American Chapter of International Group for the Psychology of Mathematics Education (PME-NA24, Vol.2, pp.539-550), Athens,GA.

Confrey, J., & Carrejo, D. (2002b). *Can high stakes testing in Texas inform instructional decision-daking?* Paper presented at the Twenty-fourth Annual Meeting of the North American Chapter of International Group for the Psychology of Mathematics Education (PME-NA24, Vol. 2, pp.551-563), Athens, GA.

Confrey, J., Castro-Filho, J., & Wilhelm, J. (2000). Implementation research as a measure to link systemic reform and applied psychology in mathematics education. *Educational Psychologist, 35*(3), 179-191.

Confrey, J., & LaChance, A. (2000). Transformative teaching experiments through conjecture-driven research design. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Confrey, J., & Makar, K. (in press). Critiquing and improving the use of data from high stakes tests: Understanding variation and distribution in relation to equity using dynamic statistics software. In C. Dede (Ed.), *Proceedings for Scaling up for success: Lessons learned from technology-based educational improvement*. Cambridge, MA: Harvard College of Education.

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

Confrey, J., Makar, K., & Nicholson, R. (2001). Teacher-to-teacher dynamic data investigation. Part of the symposium: Teachers' use of data to support mathematics implementation research, *NCTM Research Presession*. Orlando, FL.

Confrey, J., & Smith, E. (1995). Splitting, covariation, and their role in the development of exponential functions. *Journal for Research in Mathematics Education, 26*(1), 66-86.

Cook, P. J., & Ludwig, J. (1998). The burden of "acting White": Do Black adolescents disparage academic achievement? In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 375-400). Washington, D.C.: Brookings Institution Press.

Crenshaw, K. W. (1988). Race, reform, and retrenchment: Transformation and legitimation in antidiscrimination law. *Harvard Law Review, 101,* 1331-1387.

Cuban, L. (1990). Reforming again, again, and again. *Educational Researcher, 19*(1), 3-13.

delMas, R. (in press). A comparison of mathematical and statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

delMas, R., Garfield, J., & Chance, B. L. (2001). *Tools for teaching and assessing statistical inference*. University of Minnesota. Available: http://www.gen.umn.edu/faculty_staff/delmas/stat_tools/ [2001, January 6].

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).

Delpit, L. (1988). The silenced dialogue: Power and pedagogy in educating other people's children. *Harvard Educational Review, 58*(3), 280-298.

Delpit, L. (1996). *Other people's children: Cultural conflict in the classroom*. New Press.

Dewey, J. (1997). *Experience and education*. New York: Simon and Schuster. Original work published 1938.

DuBois, W. E. B. (1903). The talented tenth, *Negro problem: A series of articles by representative negroes of today*. New York: Available online at: http://www.yale.edu/glc/archive/1148.htm.

Duckworth, E. (1996). *"The having of wonderful ideas" and other essays on teaching and learning* (2nd ed.). New York: Teachers College Press.

Edelson, D. C. (2002). Design research: What we learn when we engage in design. *The Journal of the Learning Sciences, 11*(1), 105-121.

Edwards, R. (1996). *Teaching statistics: Teacher knowledge and confidence.* Paper presented at the 19th Annual Conference of the Mathematics Education Research Group of Australasia.

Evans, N., Forney, D., & Guido-DiBrito, E. (1998). *Student development in college: Theory, research, and practice*. San Francisco: Jossey-Bass.

Everitt, B. S. (1998). *The Cambridge dictionary of statistics*.: Cambridge University Press.

Fennema, E., & Franke, M. (1992). Teachers' knowledge and its impact. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 147-164). New York: MacMillan.

Finzer, W. (2001). Fathom! (Version 1.16) [Computer Software]. Emeryville, CA: KCP Technologies.

Fischbein, E., & Schnarch, D. (1997). Brief report: The evolution with age of probabilistic, intuitively based misconceptions. *Journal for Research in Mathematics Education, 28*(1), 96-105.

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

Foley, D. E. (1997). Deficit thinking models based on culture: The anthropological protest. In R. Valencia (Ed.), *The evolution of deficit thinking: Educational thought and practice* (pp. 113-131). London: Falmer Press.

Fosnot, C. T. (1996). Teachers construct constructivism: The center for constructivist teaching/teacher preparation project. In C. T. Fosnot (Ed.), *Constructivism: Theory, perspectives, and practice* (pp. 205-216). New York: Teachers College, Columbia University.

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

Gal, I. (Ed.). (in press). *Statistics Education Research Journal: Special Issue on Statistical Reasoning about Variation*. http://www.stat.auckland.ac.nz/serj.

Gal, I., & Garfield, J. (Eds.). (1997). *The assessment challenge in statistics education*. Amsterdam: IOS Press.

Gamoran, A., & Hannigan, E. C. (2000). Algebra for everyone? Benefits of college-preparatory mathematics for students with diverse abilities in early secondary school. *Educational Evaluation and Policy Analysis, 22*(3), 241-254.

Gardner, H., & Hudson, I. (1999). University students' ability to apply statistical procedures. *Journal of Statistics Education, 7*(1).

Garfield, J. (2003). Assessing statistical reasoning. *Statistical Education Research Journal, 2*(1), 22-38.

Garfield, J., Ben-Zvi, D., & Mickelson, W. (2002). The Third International Research Forum on Statistical Reasoning, Thinking, and Literacy: Second announcement (November 2002). University of Nebraska.

Garfield, J., & Burrill, G. (Eds.). (1997). *Research on the role of technology in teaching and learning statistics*. Voorburg, The Netherlands: International Statistics Institute.

Garfield, J., & Chance, B. L. (2000). Assessment in statistics education: Issues and challenges. *Mathematical Thinking and Learning, 2*(1&2), 99-126.

Garfield, J., & Gal, I. (1999). Teaching and assessing statistical reasoning. In L. Stiff & F. Curcio (Eds.), *Developing mathematical reasoning in grades K-12*. Reston VA: National Council of Teachers of Mathematics.

Glennan, T. K., & Melmed, A. (2000). Challenges of creating a nation of technology-enabled schools, *Technology and learning*. San Francisco: Jossey-Bass.

Gray, J. (2000). *Teachers at the center: A memoir of the early years of the National Writing Project*. Berkeley: The National Writing Project.

Hancock, C. (1995). Tabletop. Navato, CA: Broderbund Software Direct.

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

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

Hawkins, J. (1997). *Learning and technology: Integrating policy perspectives and research*. Boulder, CO: (Draft) Report to the Education Commission of the States, National Science Foundation.

Heaton, R., & Mickelson, W. (2002). The learning and teaching of statistical investigation in teaching and teacher education. *Journal of Mathematics Teacher Education, 5*(1), 35-59.

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

Hickman, L. A. (1990). *John Dewey's pragmatic technology*. Bloomington: Indiana University Press.

Howe, K. R. (2003). *Closing methodological divides: Toward democratic educational research* (Vol. 11). Dordrecht, Netherlands: Kluwer Academic Publishers.

International Society for Technology in Education. (2000). National education technology standards for students, *Technology and learning*. San Francisco: Jossey-Bass.

Jackiw, N. (2001). Geometer's Sketchpad (Version 4). Emeryville, CA: Key Curriculum Press.

Jencks, C., & Phillips, M. (Eds.). (1998). *The Black-White test score gap*. Washington, D.C.: The Brookings Institution.

Johnson, R. S. (2002). *Using data to close the acheivement gap: How to measure equity in our schools*. Thousand Oaks, CA: Corwin Press.

Jones, G., Langrall, C., Thornton, C., & Mogill, A. T. (1999). Students' probabilistic thinking in instruction. *Journal for Research in Mathematics Education, 30*(5), 487-519.

Kahle, J. B. (1996, April 24). *Thinking about equity in a different way.* Paper presented at the NRC: CCSSO, Seattle, WA.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press.

Kamii, C. (1994). *Young children continue to reinvent arithmetic - 3rd grade: Implications of Piaget's theory*. New York: Teachers College Press.

Kaput, J. J. (1992). Technology and mathematics education. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 515-556). Reston, Va: National Council of Teachers of Mathematics.

King, P. M., & Kitchener, K. S. (1994). *Developing reflective judgment*. San Francisco: Jossey-Bass.

Kirkpatrick, H., & Cuban, L. (2000). Should we be worried? What the research says about gender differences in access, use, attitudes, and achievement with computers, *Technology and learning* (pp. 155-167). San Francisco: Jossey-Bass.

Klein, V., & Remillard, J. T. (2002). *Navigating the worlds of conceptually-oriented mathematics and standardized test preparation.* Paper presented at the Twenty-fourth Annual Meeting of the North American Chapter of International Group for the Psychology of Mathematics Education (PME-NA24, Vol. 2, pp. 681-684), Athens, GA.

Koedinger, K. (1998). Conjecturing and argumentation in high school geometry students. In R. Lehrer & D. Chazan (Eds.), *Designing learning environments for developing understanding of geometry and space* (pp. 319-347). Mahwah, NJ: Lawrence Erlbaum Associates.

Konold, C. (2002a). *Alternatives to scatterplots.* Paper presented at the Sixth International Conference on Teaching Statistics (ICOTS6), Cape Town, South Africa.

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

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

Konold, C., Higgins, T., Russell, S. J., & Khalil, K. (2003). *Data seen through different lenses*. Unpublished manuscript.

Konold, C., & Khalil, K. (2003). *If u can graff these numbers - 2, 15, 6 - your stat literat.* Paper presented at the American Educational Research Association, Chicago.

Konold, C., & Miller, C. (2002). Tinkerplots (v. 0.45). Amherst, MA: SRRI, University of Massachusetts.

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., Robinson, A., Khalil, K., Pollatsek, A., Well, A., Wing, R., & Mayr, S. (2002). *Students' use of modal clumps to summarize data.* Paper presented at the Sixth International Conference on Teaching Statistics: Developing a Statistically Literate Society, Cape Town, South Africa.

Kuhn, T. (1996). *The structure of scientific revolutions* (3rd ed.). Chicago, Il.: University of Chicago Press. Original work published 1961.

Kurtz, M. (1999, September 19, 1999). Tutoring sets blacks apart - and gets results. *Austin American Statesman*, pp. A1.

LaChance, A. (1999). *Promoting reform in mathematics education by building content knowledge, technological skills, and teacher community.* Unpublished doctoral dissertation, Cornell University, Ithaca, New York.

LaChance, A., & Confrey, J. (2003). Interconnecting content and community: a qualitative study of secondary mathematics teachers. *Journal of Mathematics Teacher Education, 6*(2), 107-137.

Ladson-Billings, G. (1995). Making mathematics meaningful in multicultural contexts. In W. G. Secada & E. Fennema & L. B. Adajian (Eds.), *New directions for equity in mathematics education* (pp. 126-145). Cambridge: Cambridge University Press.

Lakatos, I. (1976). *Proofs and refutations: The logic of mathematical discovery*. Cambridge: Cambridge University Press.

Land, S. M., & Hannafin, M. J. (1996). A conceptual framework for the development of theories-in-action with open-ended learning environments. *Educational Technology Research and Development, 44*(3), 37-53.

Land, S. M., & Hannafin, M. J. (1997). Patterns of understanding with open-ended learning environments: A qualitative study. *Educational Technology Research and Development, 45*(2), 47-73.

Langrall, C., & Mooney, E. (2002). *The development of a framework characterizing middle school students' statistical thinking.* Paper presented at the Sixth International Conference on Teaching Statistics (ICOTS6), Cape Town.

LaTurner, R. J. (2003). *UTeach end of year report: Fall 2003*. Austin, Texas: University of Texas at Austin.

Lee, C. (Ed.). (2003). *Proceedings of the Third International Research Forum on Statistical Reasoning, Thinking, and Literacy: Reasoning about Variation*.: Central Michigan University.

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

Lehrer, R., & Schauble, L. (2000b). Inventing data structures for representational purposes: Elementary grade students' classification models. *Mathematical Thinking and Learning, 2*(1&2), 51-74.

Lehrer, R., & Schauble, L. (2002). *Distribution: A resource for understanding error and natural variation.* Paper presented at the Sixth International Conference on Teaching Statistics (ICOTS6), Cape Town.

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

Lieberman, A., & Wood, D. R. (2003). *Inside the National Writing Project*. New York: Teachers College Press.

Lomax, R. G. (2001). *An introduction to statistical concepts for education and the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Loucks-Horsley, S., Hewson, P.W., Love, N., & Stiles, K. E. (1998). *Designing professional development for teachers of science and mathematics.* Thousand Oaks, CA: Corwin Press, Inc.

Love, N. (2002). *Using data/getting results: A practical guide for school improvement in mathematics and science*. Norwood, MA: Christopher-Gordon Publishers.

Love, N. (2003). Uses and abuses of data. *ENC Focus, 10*.

Lynch, S. J. (2000). *Equity and science education reform*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ma, L. (1999). *Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States.* Mahwah, NJ: Lawrence Erlbaum Associates.

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

Makar, K., & Confrey, J. (2003). *Clumps, chunks, and spread out: Secondary preservice teachers reasoning about variation.* Paper presented at the Third International Research Forum on Statistical Reasoning, Thinking, and Literacy, Lincoln, NE.

Makar, K., & Confrey, J. (2004). *Modeling fairness in student achievement in mathematics using statistical software by preservice secondary teachers.* Paper presented at the ICMI Study 14: Applications and modeling in mathematics education, Dortmund, Germany.

Makar, K., & Confrey, J. (in press). Secondary teachers' reasoning about comparing two groups. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking*. Dordrecht, the Netherlands: Kluwer Academic Publisher.

Makar, K., & Confrey, J. (under review). Secondary preservice teachers' informal reasoning about variation. *Statistics Education Research Journal (http://www.stat.auckland.ac.nz/serj).*

Mansilla, V. B., Miller, W. C., & Gardner, H. (2000). On disciplinary lenses and interdisciplinary work. In S. Wineburg & P. Grossman (Eds.), *Interdisciplinary curriculum: Challenges to implementation*. New York: Teachers College Press.

Marshall, J., Makar, K., & Kazak, S. (2002). *Young urban students' conceptions of data uses, representation, and analysis.* Paper presented at the Twenty-fourth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education (PME-NA), Athens, GA.

McClain, K., & Cobb, P. (2001). Supporting students' ability to reason about data. *Educational Studies in Mathematics, 45,* 103-129.

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

Meletiou, M. (2003). Resource of literature on variation. *Statistical Education Research Journal, 1*(1).

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from Persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Mickelson, W., & Heaton, R. (in press). Primary teachers' statistical reasoning about data. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education, 26*(1), 20-39.

Moll, L. C., & Gonzalez, N. (2003). Engaging life: A funds of knowledge approach to multicultural education. In J. A. Banks & C. A. M. Banks (Eds.), *Handbook of research on multicultural education* (2nd ed., pp. 699-715). San Francisco: Jossey-Bass.

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review, 65*(2), 123-165.

Moritz, J. (in press). Reasoning about covariation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

National Commission on Teaching and America's Future. (1996). *What matters most: Teaching for America's future*. New York: Rockefeller Foundation and Carnegie Corporation.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics*. Reston, VA: author.

National Council of Teachers of Mathematics. (1995). *Assessment standards for school mathematics*. Reston, VA: author.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Research Council. (1996). *National science education standards*. Washington, D.C.: National Academy Press.

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

National Research Council. (2001a). *Educating teachers of science, mathematics, and technology: New practices for the new millennium*. Washington, D.C.: National Academy Press.

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

National Research Council. (2002). *Scientific research in education*. Washington D.C.: National Academy Press.

National Research Council, & Mathematical Sciences Education Board. (1989). *Everybody counts: A report to the nation on the future of mathematics education*. Washington, D.C.: National Academy Press.

National Writing Project. (2002a). *National Writing Project mission*. Author. Available: www.writingproject.org [2002, April 28].

National Writing Project. (2002b, September). *National Writing Project in brief*. University of California.
Available: http://writingproject.org/downloads/nwpinbrief.pdf [2002, October].

National Writing Project. (2002c). *Profiles of the National Writing Project*. Available: www.writingproject.org/downloads/profiles.pdf [2002, October].

Noss, R., Pozzi, S., & Hoyles, C. (1999). Touching epistemologies: Meanings of average and variation in nursing practice. *Educational Studies in Mathematics, 40*, 25-51.

Oakes, J. (1990). *Multiplying inequalities: The effects of race, class, and tracking on opportunities to learn math and science*. Santa Monica: RAND.

Ogbu, J. U. (1992). Understanding cultural diversity and learning. *Educational Researcher, 21*(8), 5-14.

Ogbu, J. U. (1994). Racial stratification and education in the United States: Why inequality persists. *Teachers College Record, 96*(2), 264-298.

Orfield, G., & DeBray, E. H. (Eds.). (1999). *Hard work for good schools: Facts, not fads in Title I reform*. Cambridge, MA: The Civil Rights Project, Harvard University.

Orfield, G., & Kornhaber, M. L. (Eds.). (2001). *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. New York: The Century Foundation Press.

Pallas, A. M., & Alexander, K. L. (1983). Sex differences in quantitative SAT performance: New evidence on the differential coursework hypothesis. *American Educational Research Journal, 20*(2), 165-182.

Papert, S. (1990). *A critique of technocentrism in thinking about the school of the future*. Available: www.papert.org/articles/ACritiqueofTechnocentrism.html [2001, January 30, 2001].

Peirce, C. S. (1998). *Chance, love, and logic: Philosophical essays*. Lincoln, NE: Bison Books. Original work published 1923.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, D.C.: National Academy Press.

Perry, W. G. (1999). *Forms of ethical and intellectual development in the college years: A scheme*. San Francisco: Jossey-Bass. Original work published 1968.

Pfannkuch, M. (2004). SRTL-4, first announcement. Auckland, NZ: University of Auckland.

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., & Wild, C. (2001). What do we know about statistical thinking? Overview of statistical thinking, a literature review. Prereading for the Second International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL2); Armidale, Australia.

Phillips, M., Brooks-Gunn, J., Duncan, G. J., Klebanov, P., & Crane, J. (1998). Family background, parenting practices, and the Black-White test score gap. In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap*. Washington, D.C.: Brookings Institution Press.

Polman, J. L. (2000). *Designing project-based science: Connecting learners through guided inquiry*. New York: Teachers College Press.

Porter, A. C., Archbald, D. A., & Tyree, A. K., Jr. (1990). Reforming the curriculum: Will empowerment policies replace control? In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing: The 1990 yearbook of the Politics of Education Association* (pp. 11-36). London: Falmer Press.

Powell, M. J. (1994). *Equity in the reform of mathematics and science education*. Austin, TX: Southwest Educational Development Laboratory.

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

Reading, C., & Shaughnessy, J. M. (in press). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Reichardt, C. S., & Gollob, H. F. (1997). When confidence intervals should be used instead of statistical significance tests, and vice versa. In L. L. Harlow & S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (Vol. I, pp. 259-284). Mahwah, NJ: Lawrence Erlbaum Associates.

Rossman, A. J., Chance, B. L., & Lock, R. H. (2001). *Workshop statistics with Fathom*. Emeryville, CA: Key College Press.

Roth, W.-M., & McGinn, M. K. (1997). Graphing: Cognitive ability or practice? *Science Education, 81*, 91-106.

Rubin, A. (2002). *Interactive visualizations of statistical relationships: What do we gain?* Paper presented at the Sixth International Conference on Teaching Statistics (ICOTS6), Cape Town.

Saldanha, L. A., & Thompson, P. W. (2001). Students' reasoning about sampling distributions and statistical inference. Prereading for the Second International Research Forum on Statistical Reasoning, Thinking, and Literacy, Armidale, Australia.

Scheaffer, R. L., Watkins, A. E., & Landwehr, J. M. (1998). What every high-school graduate should know about statistics. In S. Lajoie (Ed.), *Reflections on statistics: Learning, teaching, and assessment in grades K-12*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Scheurich, J. J., & Skrla, L. (2001). Continuing the conversation on equity and accountability: Listening appreciatively, responding responsibly. *Phi Delta Kappan*, 322-326.

Scheurich, J. J., & Skrla, L. (2003). *Leadership for equity and excellence*. Thousand Oaks, CA: Corwin Press.

Schmoker, M. (1996). Performance Data, *Results: The key to continuous school improvement* (pp. 29-48). Alexandria, Virginia: Association for Supervision and Curriculum Development (ASCD).

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

Schoenfeld, A. H. (2002). Making mathematics work for all children: Issues of standards, testing, and equity. *Educational Researcher, 31*(1), 13-25.

Schwab, J. J. (1978a). *Science, curriculum, and liberal education*. Chicago: University of Chicago Press.

Schwab, J. J. (1978b). Testing and the curriculum. In I. Westbury & N. J. Wilkof (Eds.), *Science, curriculum, and liberal education: Selected essays of Joseph J. Schwab* (pp. 275-286). Chicago: The University of Chicago Press.

Secada, W. G. (1992). Race, ethnicity, social class, language, and achievement in mathematics. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 623-660). Reston, VA: National Council of Teachers of Mathematics.

Secada, W. G. (1994). Equity in restructured schools. *NCRMSE Research Review, 3*(3), 16-20.

Secada, W. G. (2000). Foreword. In S. J. Lynch, *Equity and science education reform* (pp. ix-xvi). Mahwah, NJ: Lawrence Erlbaum Associates.

Seymour, E., & Hewitt, N. M. (1997). *Talking about leaving: Why undergraduates leave the sciences*. Boulder, CO: Westview Press.

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). Reston, VA: National Council of Teachers of Mathematics.

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

Shaughnessy, J. M., & Bergman, B. (1993). Thinking about uncertainty: Probability and statistics. In P. Wilson (Ed.), *Research ideas for the classroom: High school mathematics (NCTM)* (pp. 177-197). New York: Macmillan Publishing Company.

Shaughnessy, J. M., Watson, J., Moritz, J., & Reading, C. (1999). *School mathematics students' acknowledgement of statistical variation.* Paper presented at the NCTM Research Presession Symposium: *There's More to Life than Centers*. Paper presented at the 77th annual NCTM Conference, San Francisco.

Siegel, M., & Borasi, R. (1994). Demystifying mathematics education through inquiry. In P. Ernest (Ed.), *Constructing mathematical knowledge: Epistemology and mathematics education* (Vol. 4, pp. 201-214). Washington, D.C.: Falmer Press.

Simon, M. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education, 26*(2), 114-145.

Simon, M. (2000). Characterizing a perspective underlying the practice of mathematics teachers in transition. *Journal for Research in Mathematics Education, 31*(5), 579-601.

Smith, M. (2004). Personal communication (April 16, 2004). Austin, TX.

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

Stage, F. K., & Maple, S. A. (1996). Incompatible goals: Narratives of graduate women in the mathematics pipeline. *American Educational Research Journal, 33*(1), 23-51.

State Board for Educator Certification. (2004). Temporary teacher certificate: Guidelines for candidates. www.sbec.state.tx.us/CBECOonline/certinfo/tempcert/candinfo.asp.

Steinbring, H. (1990). The nature of statistical knowledge and the traditional mathematics curriculum: Some experience with inservice training and developing materials. In Hawkins, A. (Ed.) *Teaching teachers to teach statistics*. Voorburg, The Netherlands: International Statistics Institute.

Stigler, J., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. New York: Free Press.

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

Stuart, M. (1995). Changing the teaching of statistics. *The Statistician, 44*(1), 45-54.

Tapscott, D. (2000). The digital divide, *Technology and learning* (pp. 127-154). San Francisco: Jossey-Bass.

Tate, W. (1994). Race, retrenchment, and the reform of school mathematics. *Phi Delta Kappan*, 477-484.

Tate, W. (1995a). Economics, equity, and the national mathematics assessment: Are we creating a national toll road? In W. G. Secada & E. Fennema & L. B. Adajian

(Eds.), *New directions for equity in mathematics education* (pp. 191-206). New York: Cambridge University Press.

Tate, W. (1995b). School mathematics and African American students: Thinking seriously about opportunity-to-learn standards. *Educational Administration Quarterly, 31*(3), 424-448.

Tate, W. (2001). Science education as a civil right: Urban schools and opportunity-to-learn considerations. *Journal of Research in Science Teaching, 38*(9), 1015-1028.

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

Texas Education Agency. (1997). *Texas Essential Knowledge and Skills*. Available: http://www.tea.state.tx.us/teks/.

Texas State Board of Education. (2000). Excerpts from Long-Range Plan for Technology, 1996-2010, *Technology and learning* (pp. 38-47). San Francisco: Jossey-Bass.

Thompson, A. G. (1992). Teachers' beliefs and conceptions: A synthesis of the research. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 127-146). New York: McMillan.

Thompson, P. (2001). Conceptual issues in understanding sampling distributions and margin of error. Presented at the Second International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL2); Armidale, Australia.

Thorson, A. (2000). Editorial: Educational equity-a moving target. *ENC Focus, 7,* 4-5.

Tinker, R. (1996). *Information technologies in science and mathematics education*. Eisenhower National Clearinghouse (ENC).
Available: www.enc.org/professional/research/journal/documents/0,1341,ACQ-104633-4633_01,00.shtm [2000, October 30].

Toulmin, S. (2001). *Return to reason*. Cambridge: Harvard University Press.

Tyack, D., & Cuban, L. (1995). *Tinkering toward utopia: A century of public schools reform*. Cambridge, MA: Harvard University Press.

Tyler, R. W. (1969). *Basic principles of curriculum and instruction*. Chicago: University of Chicago Press. Original work published 1949.

U.S. Congress Office of Technology Assessment. (1988). *Power on! New tools for teaching and learning*. Washington, D.C.: U.S. Government Printing Office.

U.S. Congress Office of Technology Assessment. (1995). *Teachers and technology: Making the connection*. Washington, D.C.: U.S. Government Printing Office.

U.S. Department of Education. (2001). *No Child Left Behind-Executive summary*. Available: http://www.ed.gov/nclb/overview/intro/execsumm.html.

U.S. Department of Education. (2002). *Strategic plan 2002-2007*. Available: http://www.ed.gov/about/reports/strat/plan2002-07/plan.pdf.

U.S. Department of Education. (2003). *Educational practices supported by rigorous evidence: A user friendly guide*. Washington, D.C.: Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Utts, J. (2002). *What educated citizens should know about statistics and probability.* Paper presented at the Sixth International Conference on Teaching Statistics, Cape Town, South Africa.

Valencia, R. (Ed.). (1997). *The evolution of deficit thinking: Educational thought and practice*. Bristol, PA: The Falmer Press.

Valenzuela, A. (1999). *Subtractive schooling: U.S.-Mexican youth and the politics of caring*. Albany: State University of New York Press.

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

Voss, J., & Post, T. (1988). On the solving of ill-structured problems. In M. Chi & R. Glaser & M. Farr (Eds.), *The nature of expertise*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Warren, B., & Rosebery, A. S. (1995). Equity in the future tense: Redefining relationships among teachers, students, and science in linguistic minority classrooms. In W. G. Secada & E. Fennema & L. B. Adajian (Eds.), *New directions for equity in mathematics education* (pp. 298-328). Cambridge: Cambridge University Press.

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

Watson, J. (2002). *Creating cognitive conflict in a controlled research setting: Sampling.* Paper presented at the Sixth International Conference on Teaching Statistics, Cape Town.

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

Watson, J., & Moritz, J. B. (2000). The longitudinal development of understanding of average. *Mathematical Thinking and Learning, 2*(1&2), 11-50.

Weissglass, J. (2002, Fall). Inequity in mathematics education: Questions for educators. *The Mathematics Educator, 12*, 34-39.

Wild, C. J., & Pfannkuch, M. (1998). What is statistical thinking? In L. Pereira-Mendoza & L. Kea & T. Kee & W. Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching of Statistics (ICOTS-5) 21-26 June 1998* (Vol. 1, pp. 333-339). Voorburg, Netherlands: International Statistical Institute.

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

Wiliam, D. (2003). Constructing difference: Assessment in mathematics education. In L. Burton (Ed.), *Which way social justice in mathematics education?* (pp. 189-207). Westport, CT: Praeger.

Wilson, S. M., & Berne, J. (1999). Teacher learning and the acquisition of professional knowledge: An examination of research on contemporary professional development. *Review of Research in Education, 24*, 173-209.

Zeverbergen, R. (2003). Teachers' beliefs about teaching mathematics to students from socially disadvantaged backgrounds: Implications for social justice. In L. Burton (Ed.), *Which way social justice in mathematics education?* (pp. 133-151). Westport, CT: Praeger.

VITA

Katie M. Makar was born Katherine Margaret Roberts in Riverside, California on March 7, 1963, the daughter of Marilyn Louise Roberts and the late James Richard Roberts. After graduating from Lake Oswego High School in Lake Oswego, Oregon in 1981, she earned her Bachelor of Arts degree in Mathematics (with Honors) from Mills College in Oakland California in 1984. She earned her California State Single-Subject (Mathematics) Teaching Credential from Mills College the same year. Following college graduation, she was employed as a mathematics teacher at Henry M. Gunn High School (Palo Alto, California), and then the Catlin Gabel School (Portland, Oregon). She completed a Master of Arts degree in Mathematics (Logic) in 1991 from the University of California (Berkeley). Following this, she taught high school mathematics for nine years overseas at the Lincoln School (Kathmandu, Nepal) and the International School of Kuala Lumpur (Kuala Lumpur, Malaysia), where she was also the department chair. In 2000, she was admitted to the mathematics education program at the University of Texas at Austin, where she worked under Dr. Jere Confrey. She has co-authored three peer-reviewed publications as well as presented her work (with Dr. Confrey) on four continents. She lives with her husband and daughter all over the world.

Permanent Address: 3745 Division Court, Lake Oswego, Oregon 97035

This dissertation was typed by the author.