Undertaking Data Analysis of Student Outcomes as Professional Development for Teachers

Jere Confrey, St. Louis (USA)
Katie Makar, Austin (USA)
Sibel Kazak, St. Louis (USA)

Abstract: The study reports on collaborations with practitioners to examine the results of students' performances on high stakes tests as a means to strengthen practitioners' knowledge of probability and statistics and to empower their conduct of investigations on student performance. Four issues are summarized: the development of their statistical reasoning, their understanding of the meaning of and relationships among the concepts of validity, reliability and fairness as applied to testing, their introduction to the history of testing and its relationship to science, society and cultural inequality, and their reports of independent inquiries. Data on performance on pre- and post-tests demonstrate growth in teacher reasoning and in their professionalism in raising important issues about testing.

Kurzreferat: *Professionalisierung von Lehrer(inne)n durch Analyse von Schüler(innen)daten*. Der Beitrag berichtet von einer Zusammenarbeit mit Praktiker(inne)n, in der durch die Analyse der Ergebnisse von Tests zu Schüler(innen)leistungen das Wissens von Praktiker(inne)n im Bereich der Wahrscheinlichkeitsrechnung und Statistik gefestigt und die Praktiker(innen) zu einem angemessenen Umgang mit Untersuchungen von Schüler(innen)leistungen befähigt werden sollen. Vier Aspekte werden dargestellt: ihre Entwicklung von statistischer Argumentation, von Verständnis von Bedeutung und Beziehungen zwischen den Konzepten Gültigkeit, Zuverlässigkeit und Fairness beim Testen; ihre Einführung in die Geschichte des Testens und dessen Verhältnis zu Wissenschaft, Gesellschaft und kultureller Ungleichheit, und Berichte von unabhängigen Untersuchungen. Ergebnisse von Pre- und Posttests weisen auf eine positive Entwicklung in der Argumentation der Lehrer(innen) und in ihrer Professionalität relevante Aspekte des Testens zu erkennen hin.

ZDM-Classification: B55, C39, D39, D60, K49

1. Introduction

In the United States, there is a need for mathematics teachers to become more knowledgeable about reasoning with data using statistics. Increasingly, statistics and data analysis are common parts of K-12 curricula, and represent a key shift in intellectual thought, permitting citizens and professionals to examine numerous complex phenomena of social importance. Furthermore, the inclusion of data and statistical reasoning into mathematics classes signals a major epistemological shift in mathematics education – towards issues of uncertainty, approximation, modeling, estimation and prediction in relation to context, and away from nearly exclusive focus on deduction, proof, definition and abstract mathematical systems. With the advent of increasingly powerful technologies for acquiring, storing, retrieving, analyzing and displaying data, the study of statistics brings with it a greater reliance on application of reasoning to complex problems, the thoughtful choice of procedures, approaches, and the amalgamation of evidence, persuasion, and prediction. Together this array of factors make it imperative to help mathematics teachers gain proficiency in teaching data and statistics, by strengthening their own knowledge and shifting their mathematical perspective.

Simultaneously, due to federal legislation entitled the *No Child Left Behind Act*, teachers and schools in the United States are subject to intense pressures from high stakes tests and accountability systems that tie student progress, financial consequences, and public scrutiny to their performance on annual tests. Many professional educators are skeptical about the movement, fearing that it focuses unduly on a narrow set of heavily procedural skills, and tends to encourage teaching to the test, restricting broad curricular preparation. Typically, schools, districts and teachers are provided data from the tests in the form of summary statistics, such as percent passing or mean scores, and these are increasingly disaggregated by subgroup performance (race, gender, socio-economic class, special education designation, and language proficiency). Some tests are also disaggregated by content strand score, however, typically, the whole test is scaled for test equivalence from year to year, leaving the validity of the subscores by content strand in question (Confrey & Carrejo 2002). There is widespread dispute over whether these approaches exacerbate inequalities or help remedy them. Both sides acknowledge that the tests document widespread gaps in performance, particularly in relation to poverty and certain at-risk ethnic groups (particularly for Hispanics and African Americans), and some would claim that by directing attention to these gaps, improvements are more likely to follow. Others counter that the intensive testing and accountability put undue burdens on children, exacerbate the impact of disparities in educational resources, and widen rather than lessen educational opportunity (McNeil & Valenzuela 2001).

As a research group who works closely with urban schools, we have found ourselves caught in the competing trends of a movement to include increased emphasis on high cognitive thinking that is coupled to a demand for documented progress as defined largely by multiple choice tests. We have named this tension *systemic crossfire* (Confrey, Bell & Carrejo 2001). For instance, for five years in partnership with an urban high school, we had been able to document steady gains on the tests as teachers implemented increased uses of technology and standards-based curriculum material. In addition, researchers documented improved teacher community (Lachance 1999; Castro-Filho 2000). During one single year, a relatively small subgroup of African American students (n = 31) dropped below acceptable passing levels (below 50%) rendering the school "low-performing". Despite our demonstrating that the performance of the subgroup was not unexpected – as a chance dip in performance within a trajectory of improved performance – existing state mandates, ensuing school leadership decisions and teacher choices led to a year of school-wide test preparation and a dismantling of all our work at the school (Confrey & Makar in press; Lachance & Confrey 2003; Confrey, Castro-Filho & Wilhelm 2000). We de-

parted, realizing a need to reconfigure our relationship to the system. The failed partnership led us to accelerate and refocus an approach that we had begun at the middle grades the previous year working with teachers and using data from high stakes tests (Confrey & Makar 2002). We report on the extensions of those efforts in this paper.

As we reconsidered our model in relation to high stakes testing and accountability, we shifted to the question of how to interpret the results of these tests and to examine them for the implications with regard to equity and subsequent instructional decision-making. Our idea was to imitate the very successful national program for teaching writing, in which researchers took the position that quality instruction in writing is best achieved through immersing teachers in the activity of writing as writers (Lieberman & Wood 2003; Gray 2000). Although we did not see a simple analogy with the writing program for improving mathematics instruction through immersing teachers in the doing of creative mathematics, we did see a potential path for improving statistical instruction by having teachers conduct studies of student performance, which would immerse them in statistics and provide authentic engagement in data analysis. Our assumption was that by doing so, we would see improvement in their instruction on data and statistics, and at the same time strengthen their professional position as arbitrators of the information and pressures from the high stakes tests.

Over the past year, we conducted two research studies on statistics understanding and reasoning of practitioners. One study was conducted as part of ongoing dissertation research on pre-service teachers in Texas in spring 2003 (Makar in preparation). The other was conducted as research on practitioners in advanced study in education in Missouri in fall 2003. We report on common elements of both studies.

In each study, we developed and taught a one-semester course in professional development on assessment in which pre-service teachers, teachers in continuing education program, and graduate students learned about high stakes testing and undertook studies using real data sets using a statistical software tool called Fathom Dynamic Statistics (Finzer 2001). The courses shared many common features which included:

1) An introduction to key technical concepts of testing
2) Access to and use of data on student performance by individual schools
3) Training and use of Fathom, a dynamic statistics software environment particularly suited to new learners
4) The development of concepts of central tendency, spread, correlation, linear regression, sampling variability, and hypothesis testing (either through simulations or using t-tests and confidence intervals)
5) Access to literature on equity and testing, especially in relation to minority performance, and
6) The assignment of an independent investigation of data in relation to issues of equity.

In both cases, instruction in the use of the software and the development of the statistical reasoning was woven into the overall instruction on assessment for about an hour a week during the first three quarters of the course. The last quarter was devoted to independent or group-designed data investigations. In both courses, there were three fundamental questions that permeated discussions during the course:

- To what extent and in what ways can an analysis of high stakes test data inform instructional decision-making? (Use of information)
- How does one decide if the outcomes between two groups are statistically and educationally significant? (Statistical reasoning) and
- Do high stakes testing systems support increased levels of equity? (Fairness)

Returning to our basic premise, we believed that careful and guided inquiry into these questions would constitute a form of professionalization of educators that would contribute to their understanding of the operation of educational systems and to the use of empirical data and statistical reasoning to guide their views of the enterprise of testing. We were neither advocates nor opponents of testing, but argued rather for the ability to discern and discover the strengths, weaknesses, challenges and opportunities in the use of testing.

In this report on these initiatives, we summarize our activities in four areas:

1) Issues in the development of the statistical reasoning of educational practitioners;
2) Understanding of the meaning of and relationships among the concepts of validity, reliability and fairness as applied to testing;
3) The history of testing and its relationship to science, society and cultural inequality, and
4) Reports on dependent inquiries conducted by our educational practitioners.

We conclude by tying these four areas together in a more general discussion of equity issues.

2. Issues in the statistical reasoning of educational practitioners

The pathways for educational practitioners to learn statistical concepts do not differ markedly from those of other novice students, but we did find that conducting such studies in the context of data about testing intensified teachers' interests. Some of our students had taken courses in statistics, but reported significant differences in their experience in the context of our courses. They report that previous studies were unduly abstract, heavily procedural, and weakly motivated.

Conceptually, there were three key areas of intellectual growth in which we witnessed significant development of deeper understanding: 1) the meaning and relevance of distributions of scores, 2) the relationships among covariance, correlation and linear regression, and 3) the role of probability in comparing the performance of two groups. We briefly discuss each.

2.1 Distributions of Scores

Typically, scores on high stakes tests are only reported as summary statistics. Either the mean score or percent passing is reported. This is seriously problematic for the consideration of issues of equity, as particular groups of students can be systematically neglected and made invisible. Reporting the means or percent passing disaggregated by

subgroup does help to draw attention to such inequities, but it can contribute to stereotyping, as people tend to believe that all or even most members of that population conform to the given measure of central tendency. Fundamentally, to overcome the weakness of this kind of reporting, one needs to understand the idea of variation in scores and begin to systematically compare outcomes.

To initially motivate these ideas visually, we began by presenting the scores of student populations from two schools, one large and one small, on the same test and asking which set of students showed higher achievement. Drawing from research by McClain and Cobb (2001), we chose to use unequal groups with disparate distributions and the same mean scores to stimulate a discussion of distribution. In particular, we noted that two student approaches emerged and competed. One was to break the groups into equal numbers and compare the resulting intervals, while the other was to break the group into equal intervals and compare the resulting numbers in the groups. Interestingly the first approach leads towards box plots while the second supports the use of histograms. Figure 1 shows the original dot plots and the two competing displays.

In box plots, the central tendency is represented by the median as four groups are created by quartiles. We believe that too often box plots are inadequately developed, which undercuts their value in a) helping people use multiplicative reasoning in analyzing comparisons and b) developing a conceptually strong interpretation of percentile rankings as used by testing companies.

At the same time, it is essential to develop the approach using histograms and means to lead to the concept of standard deviation as a statistical measure of variability in distributions and highlight issues of sample size. Our approach to standard deviation had some unique features. Like others, we developed one version of standard deviation as the square root of the sum of the squares of the distances of values from the mean, divided by n-1 (one less than the sample size) and contrasted this to mean absolute deviation (MAD) – calculating the average of the absolute deviations of values from the overall mean, which is rather an intuitive way to think of measure of variability in the data. But, in the fall course, we also linked the standard deviation to the inflection point in a normal distribution as a defining characteristic of that curve. We did so by transforming histograms into density curves. We discussed how changing the vertical axis to a percentage correct rather than a score does not alter the histogram's shape but does produce a display in which the total area of the bars is equal to one. We combined this with a discussion of shapes and distributions (i.e. skewed, uniform, and normal). We discussed the concept of a normal curve in this setting, as a symmetric distribution with inflection points and tails. This approach sets up the transition to a distribution interpreted as a probability of outcomes.

In our pre- and post- tests in both courses, we included questions to see if students understood the ideas of variation. Results were mixed, but showed overall growth. One question, for example, asked students to write at least three conclusions comparing the performance of Hispanic students with that of African American students in the context of high-stakes assessment data presented as box plots of scores for each student subgroup and a table which presents disaggregated descriptive statistics, such as the sample size, mean score, and percent passing on the test for each sample. Each conclusion was coded into one of fifteen categories of responses, then scored according to a rubric which measured the statistical complexity of a response on a scale of zero (no response) to five (distribution perspective).

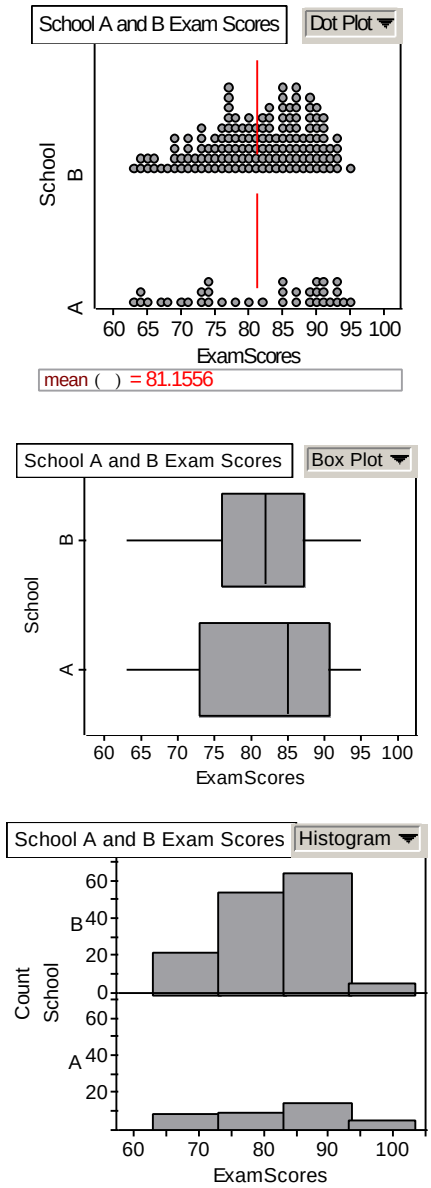

Fig. 1: Test-score data in School A and School B with the same mean and different sample sizes.

The pretest scores on the question were higher in the fall course (6.3 points) than in the spring (5.3 points); both courses showed a mean gain of about 2.6 points from pretest to posttest, with four of the 30 combined students (all above average for this question on the pretest) posting no gain. Students who began the course with lower responses on the pretest made the most gains (see Figure 2a).

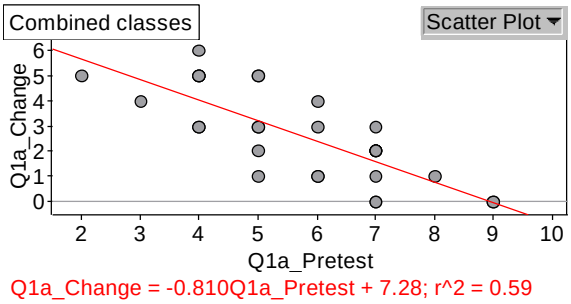

Fig. 2a: Student growth in comparing two distributions of box plots.

The overall distribution of combined student responses in both courses is displayed in Figure 2b, and can be summarized as follows:

1) In the pre-test, of thirty students, seven stated very general conclusions, while most others focused highly on percent passing or measures of central tendency (i.e. mean and median scores), neglecting variation.
2) Only a few students used some kind of reasoning about distribution in a visual sense besides centers in the pre-test.
3) None utilized the box plots in order to compare the quartiles at the beginning of the semester.
4) Even though most of the students continued comparing the two distributions with the measures of center in the post-test, about 63% of them also compared the variability between the distributions.
5) Moreover, after the course, many students (30%) were able to compare box plots of distributions, looking at the variability in quartiles and in the inter quartile range (IQR).
6) Overall, student responses in the post-test were more complete, and the emphasis was on measures additional to central tendency or percent passing.

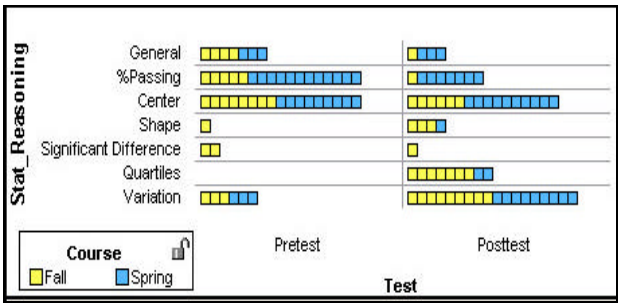

Fig. 2b: Student responses to comparing two box plots.

Finally, final projects indicated that students in the courses had developed a keener appreciation of variation and distribution of scores, including attention to missing data points or decisions to eliminate values. This is of particular importance in assessing student outcomes, for which absences or invalidating scores have been documented as a means of raising percent passing.

2.2 Covariation, Correlation, and Linear Regression

In the second statistical topic, we examined the issues of covariation, where we use this term to refer to the relationship between two variables. Adapting the treatments in Rossman, Chance, and Lock (2001) and Erickson (2000), we sought to differentiate ideas of strength and direction of the relationship through the examination of a set of exemplars. We worked with the representations to show students how these two dimensions could be linked into a single scale from -1 to 1 in which zero would represent no strength and no direction using the relationship in testing for our contexts.

On the fall post-test only, we included two items to see whether students understood the ideas related to the linear regression and the correlation between two variables. For instance, we asked students to estimate the correlation coefficient using the information provided in a scatter plot along with the linear regression line on the graph, the equation of the linear regression with negative slope, and the r-square value. Of all students, 69% estimated the correlation coefficient calculating the square root of given r-square value and taking the negative relationship between the variables into account. The rest, however, simply tried to guess from the scatter plot looking at the direction and the strength of the relationship. We also found that one of these guesses mistakenly violated a basic property of the correlation coefficient (r can be only between -1 and 1, inclusively) and did not consider the negative association. In the other multiple-choice-type item, all students were able to choose a correct interpretation of a linear regression equation in the context of the relationship between grade point averages and standardized test scores.

2.3 Sampling Distributions

The final statistical reasoning topic was to discuss sampling distributions and confidence intervals, and use these to develop the idea of inferential statistics using software simulations or the t-test. Our students used the tutorial in Fathom in the context of voting, where they could control the likelihood that a particular outcome of a vote was "Yes" or "No". After calculating the proportion of "Yes" votes for a random sample of 100 votes, they automated drawing repeated samples (called "collect measures" in Fathom) and plotted the distributions of proportion of cases that voted "Yes" (Figure 3, with "true" probability 0.6). Through this they could see that although the "true" proportion of votes was fixed, there was a great deal of variation in outcomes due to sampling variability. After examining this simulation for proportions, we repeated a parallel exercise to predict the population mean on a math test using various sample sizes.

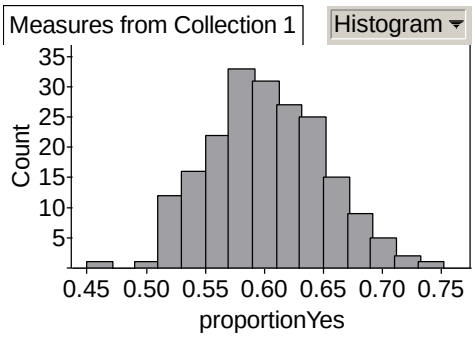

Fig. 3: Sampling distribution in Fathom displaying proportion of "Yes" votes in each sample.

For homework, we assigned students an exploration with two questions: What is the impact of a larger sample of a population on the shape of the sampling distribution? What is the impact of taking more samples on the shape of the sampling distribution? While the former exploration reveals that the larger the sample size is, the narrower the shape of the sampling distribution is (i.e. less standard error), the latter simulation shows that collecting more samples makes the shape of the sampling distribution more smooth and normal. It was clear that it is important to separate these two ideas, which are easily conflated, and to develop a strong intuition about standard error as reported in testing.

Because of our earlier work with the normal distribution, the movement to confidence intervals as a procedure for estimating a population mean was relatively straightforward. We basically followed this chain of reasoning: Based on the investigation with normal distributions, the probability is about 0.95 that the sample mean will fall within two-standard deviations of the population mean. Equivalently the population mean is likely within two-standard deviation of the sample mean, and thus 95% of all samples will capture the true population mean within two-standard deviations of the sample mean, or, if we repeat the procedure over and over for many samples, in the long run 95% of the intervals would contain the population mean. Later, the movement to the t-test as reasoning on a sampling distribution using the difference of the scores seemed relatively straightforward to our students. They used it repeatedly in their projects and through repetition or application their use of it became secure, although we do not know about their conceptual depth. In the post-test in the fall, when asked to write a conclusion statement about the given t-test output in the context of 10^{th} grade students' science test scores by gender, 85% of our students correctly responded to the question. Further, in their final project papers many reported probabilities near $p>.05$ as close, while others adjusted the level of probability in consultation with statistical consultants for valid reasons.

3. Understanding of the meaning of and relationships among the concepts of validity, reliability, and fairness as applied to testing

We used the National Research Council's (NRC) book, *High Stakes: Testing for Tracking, Promotion and Graduation* (Heubert & Hauser 1999), as a text to introduce students to the technical literature on testing. This book provided a number of resources to the course. It introduces and sets the context for the multiple uses of testing and warned of the trade offs among them. It provides a short history of the legal cases, and establishes the grounds on which challenges to the use of testing had been most prevalent. It introduces the concepts of validity, reliability and fairness. And it provides more extensive discussion of multiple settings in which testing has been examined. While the concepts of reliability were introduced, the majority of the time was spent on issues of validity and fairness.

The area that drew most debate and discussion concerned the question of how to determine when disparate impact on certain groups was not viewed as discrimination or bias in testing. As stated in the reading, "... unequal outcomes among groups do not in themselves signify test unfairness: tests may validly document group differences that are real and may be reflective in part of unequal opportunity to learn" (Heubert & Hauser 1999, p. 79). While our students could accept that a test may only be measuring discrepancies produced by a system of education, they were concerned that the consequences of this documentation of disparate impact within a system of accountability often did not lead to effective remedies. If validity was to be not simply a characteristic of a test but rather "the inferences derived from the test scores and the actions that follow" (p.73), then there appeared to be an inherent tension between fairness and equity. The tests were producing unsatisfactory results for many students of color and of poverty in terms of retention and the denial of diplomas, and yet, if constructed according to the technical specifications of experts, that result was viewed as fair. This tension seemed particularly exacerbated by the use of a single test to produce a single score, especially in light of the clear warning in the book,

> "High stakes decisions such as tracking, promotion and graduation should not automatically be made on the basis of a single test score, but should be buttressed by other relevant information about the student's knowledge and skills, such as grades, teacher recommendations, and extenuating circumstances" (p. 279).

Unevenness in the school resources or quality of instruction was not considered as a valid challenge to the legality of test results. Three questions reverberated: Is the test only a neutral monitor of a surrounding system like a thermometer is a non-intrusive device to measure the symptom of temperature, or does it change the character of instruction in intrusive ways? How much burden of the system should rest on the shoulders of individual children? And, to what extent does using a single test measurement for multiple purposes (documenting individual performance, monitoring the whole system) compromise the quality and validity of the measures? By entering into the professional, technical and legal aspects of the testing debate, informed by research, the student-practitioners began to take the dilemmas and challenges of testing more seriously, and to consider how this information shaped their own emerging opinions.

4. History of testing and its relationship to science, society and cultural inequality

The fall course began with reading *The Big Test: the Secret History of the American Meritocracy* (Lemann 2000), which tracks the development of the use of the Scholastic Aptitude Test (SAT) and its impact on American education. The book documents the SAT's inception as a means to select elite Ivy League entrants from among public school applicants in the 1920s and the evolution of its use for entry to the majority of colleges and universities. It traces the rise of a "science of intelligence testing": its initial link to eugenics and the subsequent rationale for using testing to identify intellectual elites for

selective service deferments to ensure a pool of talent for weapons and other scientific research during the Cold War and Korean War. These were tied together under the concept of a meritocracy, the primary assumption of which is that it is in the nation's interest to identify and prepare an intellectual elite motivated to serve – to lead and guide the country. The notion was made somewhat palatable to the public because many early proponents hoped to diminish family influence and economic class and replace these with merit-based "objective performance" as the primary criterion. An acknowledged weakness of the approach, however, was that it selects for "aptitude" or talent, without finding a means to select for the virtues of service and leadership.

The perennial problem with the test was that it led to differential performance that could not be assuredly linked to "intelligence." When this problem could not be solved, Educational Testing Service (ETS) switched to using the term "aptitude," which literally referred to likelihood of success in college (though the test could actually predict only 15% of the variance in first semester grades). Lemann documents how patterns of disparate performance led psychometricians to adjust the test for rural males by adding a practical science section, suppress information about Southerners' performance in relation to draft deferments, and to refocus attention on family rather than on discrepant school resources, as in the Coleman report. However, none of these cases was effectively used to challenge the validity of the test. In the case of the enormous performance discrepancies by African Americans, for example, the country responded by accepting the test differences, but instituting busing and affirmative action as means to address inequality. The discrepant performances and responses to them resulted in consistent criticism of the test's fairness, including renunciation of it by Brigham, its original creator.

Documenting how the vision of testing-linked meritocracy was shared by proponents of testing and influential leaders of elite universities, private foundations, and corporations, Lemann shows how ETS was formed, supported by the College Board, and how it gained nonprofit status. This permitted it to avoid taxes while competing for lucrative government contracts. He shows how ETS competed for adoption in the state universities with the ACT (American College Test, more oriented to achievement), and tied its tests to a vision for universal access to advanced education in California. ETS was also able to become responsible not only for the conduct and release of the research but also for its own evaluation. Ironically, this self-promoting organizational arrangement contradicts the very tenets of science that ETS used so effectively to further its own cause.

Class discussion of the *Big Test* revolved around whether the concept of a meritocracy was flawed, especially when the intellectual elite who benefit from the training are, in the main, disinclined towards service. We also debated the reasons for persistent gaps in group performance, and why these were tolerated or obscured, even as testing proponents, through political and organizational acuity, used merit as the motive and science as the means of validation to weave testing into the fabric of our culture.

5. Types of inquiries conducted by educational practitioners and their strengths and limitations

Our earlier work with middle school mathematics teachers (Confrey & Makar 2002) had shown us that teachers often begin inquiries struggling with the same issues as their students: moving from a focus on individual cases to trends, developing measurable conjectures, separating anecdote from evidence, and linking conjecture to evidence to conclusion. Our experience had taught us that after weeks of immersion in conducting data investigations into student performance, the teachers' conjectures evolved from answering simple school-level questions (e.g. On which test objective did students score the lowest?) to ones that made finer distinctions and wrestled with more complex issues and ill-structured problems (e.g. Are the school wide practice tests for the state exam beneficial – are they good predictors of actual test performance and are students taking them seriously, particularly those most at risk of failure?).

The final projects in the current study – independent inquiries into an issue of equity through an investigation of high-stakes assessment data – took place at the end of each course. In the spring, students worked independently or in pairs and had access to several data sources: recent and previous test scores and demographic backgrounds from a random sample of 10,000 students across Texas, hundreds of variables from every school in the state over the past five years available to download or display on the web from the state education agency, or data they found on the internet (e.g., from the National Center for Educational Statistics).

The spring course produced the following categories of inquiry topics:

- how race or class interacted with differences in resource allocation, achievement, or issues of disparate impact (4 projects);
- the sensitivity of the accountability system to subtle changes in school circumstances, like small subgroup size or performance variability, or differing definitions of dropout (4 projects);
- comparison studies of resources, school characteristics, and student performance between different community types, accountability ratings, or school structures—urban vs. suburban districts, schools with ratings of low-performing vs. exemplary, or magnet vs. non-magnet campuses (4 projects); and
- school case studies or current funding legislation (3 projects).

The structure of the inquiries and available data sets changed slightly in the fall course where inquiries began as group efforts and were then reported individually by the group members focusing on different aspects of a general research question. Student-level data from nine St. Louis area public schools in the same district on the Missouri Achievement Program (MAP) test on mathematics, science, and communication arts in years from 1999 to 2003 were provided for the purpose of group projects. In this setting, there were four main groups interested in: 1) investigating how racial/ethnical backgrounds, mobility, testing accommodation, and low socioeconomic status affect special-needs students' MAP test

scores on communication arts, mathematics, and science areas; 2) examining variations in student achievement, particularly in science, among demographically similar schools in a single district and identifying possible student-, teacher-, and school-level attributes that are correlated with student achievement on the MAP test; 3) studying disparities in mathematics and communication arts scores on the MAP test between the students identified as gifted and the other students and the problems of equity and efficacy in gifted education; and 4) examining the alignment of accountability system in the Missouri elementary and secondary education with the current *No Child Left Behind* legislation, predicting Missouri's projected level of future compliance using statewide and local MAP data, and determining the trends in student achievement on MAP test by disaggregated subgroups.

Thus, the variety of research foci is reflected on the range of types of inquiries used by educational practitioners and demonstrated similar type of complexity that we had seen in our earlier work with middle school teachers. Some examples will be briefly discussed next.

One group in the fall carefully examined the variation in science achievement on the MAP test among nine elementary schools in the same district. The initial descriptive analysis of the data indicated that the variation in student proficiency among these schools ranged from 17% to 60%. With this student's current statistics knowledge, the group decided to run ANOVA in order to investigate this variation further. The ANOVA result suggested that there was a statistically significant difference on students' science MAP scores among schools at the significance level of 0.05. In an attempt to explain this variation among demographically similar schools, several possible factors of which the effects on student academic achievement were suggested in the literature were investigated by correlation analysis. Specifically, student, teacher, and school related attributes (i.e. % AA, % White, % free-reduced lunch (SES indicator), % females, % males, % Limited English Proficiency (LEP), attendance rate, % satisfactory reading, % teachers with advanced degrees, school size, student to teacher ratio, years of teacher experience) were taken into account in this part of the inquiry. Then, these indicator variables were correlated with disaggregated mean scores (by gender, AA, non special education, free-reduced lunch, and non free-reduced lunch) and overall mean scores of each school. The only significant factors were found to be the student to teacher ratio, the teacher experience, and the reading proficiency as they were correlated negatively with male mean scores, positively with overall, AA, female, male, and non special education students' mean scores, and positively with overall, female, male, non special education, and non free-reduced lunch students' mean scores, respectively ($p<0.10$ was reported in all cases). One possible explanation of this result could be that the sample data were too limited in terms of variability and size to see significant correlations among other variables.

In a second example, a student in the spring course used simulations in a theoretical treatise which examined the impact of student performance variability and several remediation strategies on a school's probability of being rated low-performing by the State. In Texas, a school is labeled low-performing if the performance of its White, African-American, Hispanic, or Economically Disadvantaged subgroup (with at least 30 students and making up at least 10% of the population being tested) is below 50% passing. In her inquiry, she demonstrated how sensitive the over-simplistic measure of percent passing is (used in Texas to determine how a school is rated) to subtle changes in school situations. For example, in one simulation, she investigated what would happen to the probability that a small subgroup ($n = 32$) would fall below 50% passing if she varied the standard deviation of their scores, but kept the mean near the passing standard of 70. She found that the risk of a school going low-performing changed dramatically from near zero to over 40% when she varied their standard deviation from 1 to 30 with almost all of the change occurring for standard deviations between one and ten. In another simulation, she showed that by improving the scores of just a few students near the passing standard, known as "bubble kids" in the U.S., a school can dramatically reduce its risk of being designated low-performing from 27% to less than 2%. She emphasized through these and other simulations how the practice of rating campuses based on the measure of percent passing encourages schools to undergo questionable remediation strategies, such as focusing remediation on just a handful of students with past scores close to the passing standard or neglecting their students who are most at risk of dropping out.

The last example from students' inquiries took quite a different approach to the investigation of high-stakes data. In particular, one student attempted to use a combination of policy analysis and critical discourse analysis to investigate how the Missouri state accountability process has been documented and the way the information has been communicated with the public (parents, teachers, administrators, researchers, and concerned citizens). The primary sources of this particular inquiry were policy documents, state and federal laws, state regulations, and news releases to the public. While the use of policy analysis was reflected in student's inquiry into looking at the patterns, contradictions, and actions by the state in response to the federal legislation *No Child Left Behind*, the critical discourse analysis was nicely used as a means of looking at texts, discourses, and social practices in the public announcements in an attempt to recognize language as a form of social practice. These analyses then addressed some of the language practices in policy that are misleading and contradictory.

6. Discussion and Conclusions

In summarizing this work, which is ongoing, we can only report on the impact of the courses themselves. To date, we cannot ascertain if enrollment in the courses and engagement with the material as statisticians and analysts would improve and deepen teachers' treatment of data and statistics in their own teaching. In terms of the courses, we can report the following overall results on pre- and posttests. The pre- and posttests in both courses indicate overall gains in statistical concepts tested. The

spring pre/post test contained 26 questions that measured statistical concepts, emphasizing concepts of distribution and variation. There were significant gains in improvement from the pretest (47% average) to posttest (69% average) with an average gain of 18% - 25% at the 95% confidence level.

As might be expected, those who entered the course with higher pretest scores also tended to leave with higher posttest scores (see Figure 4, r = .67), but there was little association between pretest scores and gains made from pretest to posttest (r = -.28). This is encouraging in that it meant for those students who had previously studied statistics, they were still able to improve their statistical understanding during the course. At the same time, students who struggled at the beginning had equal opportunity to improve their understanding.

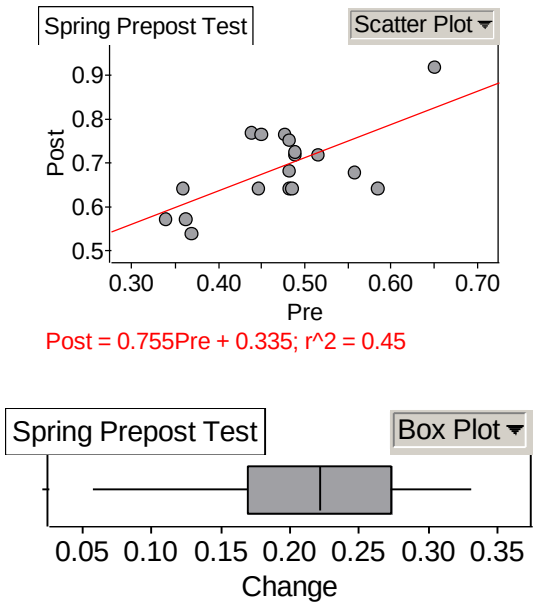

Fig. 4: Scatterplot of spring posttest versus pretest percent correct and box plot showing change in pre/post test over the spring course with 18 students.

In the fall course pre/post-test analysis, we looked at students' performances on four statistics items (see Figure 5). The box plots for the distributions of scores on pre- and posttests revealed that the majority of the students performed better after the course treatment. There was less variation in the middle 50% of the scores as well. There seemed to be a mean difference in favor of after-treatment-effect on the students' performances. The second box plot representation shows the distribution of change in scores over the course. The shape of the distribution is left-skewed (the mean is less than the median) and the top 75% of the distribution indicates gains in scores after the course. More specifically, most of the gains in scores came from the item on an important property of the measures of center and variability, namely the concept of resistance to the outliers.

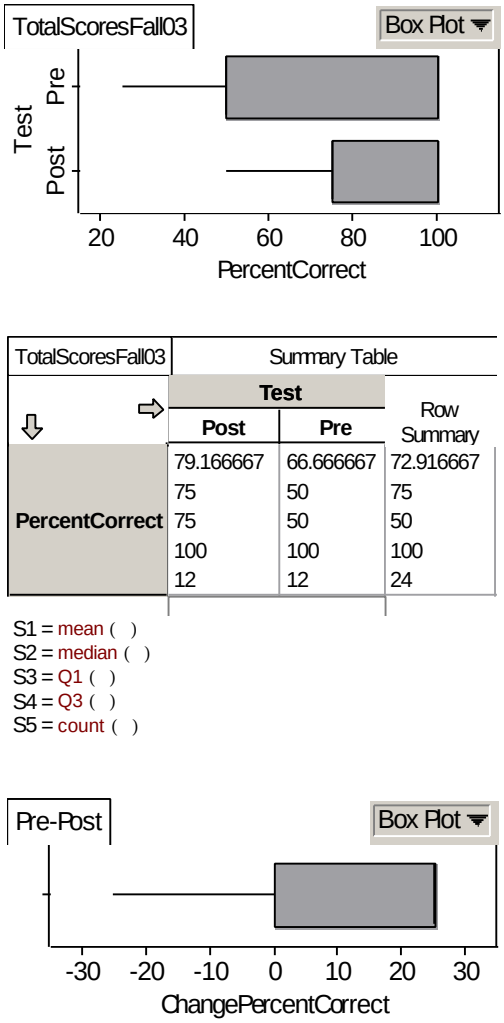

Fig. 5: Distributions of percent correct in pre- and posttests with summary statistics and change in percent correct over the fall course.

Our experience with the course convinces us of the value of involving practitioners directly in the examination and analysis of data. We were convinced that it was valuable for practitioners to recognize the importance of examining distributions to understand the dangers of summary data, to learn that differences in mean performances or percent passing is only a partial telling of the data, and to consider probability and sampling distributions in comparing groups. As they begin their teaching careers, this knowledge could urge them beyond simplistic judgment towards an enriched lens on ways that they, and their colleagues, interpret their students' test scores. The particular interaction of statistical content knowledge, use of dynamic technology, experience in conducting inquiries as learners, context of testing, and emphasis on examining issues relevant to equity we believe to be mutually supportive. We have seen that the relevance of the context for these teachers supports their understanding and motivation to learn the statistical content, which in turn allows them to dig further into their understanding about equity and testing. Likewise, the experience in data analysis provided them with ways to strengthen chains of reasoning in issues that were otherwise sensitive to discuss.

We were also convinced of the value of providing these practitioners with more understanding of the technical aspects of testing, and to place this information into legal, historical and political contexts. Their compelling and competent choices of investigations show that this audience was able to examine a raw data set and to conduct independent inquiries.

While we initially viewed the pursuit of this approach as a means to achieve better mathematics instruction in data analysis and statistics, we found rather that it was a very powerful means to reconceptualize our relationship to our professional colleagues in schools. Engaging with our prospective teachers and practitioners has led us to consider more deeply our own professional obligations, and to begin to articulate how it is that equity and excellence must not be in conflict, but mutually reinforcing. Our examination of testing in the company of practitioners also convinces us that it is unsatisfactory to have a schism between the communities who engage in the daily teaching of students and those that design, build, research and evaluate tests. It is not enough for expert practitioners to review test standards, to submit or review items for bias, and be handed summary statistics of the performance of their students, themselves and the school and district. They must be able to understand and critically examine the process of test construction, measurement and sampling assumptions, and methods of analysis such as item response theory (IRT), scoring, scaling, and test equating. Access to raw data must be easy and complete, so that they can launch their own inquiries and draw their own conclusions. Alternative test scores must become available to permit practitioners and stakeholders to challenge test validity and fairness. Only then we will truly have accountability systems that demand and permit responsible participation by practitioners and add to the overall professionalization of the field.

References

Castro-Filho, J. A. (2000): Teachers, math, and reform: An investigation of learning in practice. - University of Texas at Austin, Doctoral Diss.

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. (2002): A content analysis of exit level mathematics on the Texas Assessment of Academic Skills: Addressing the issue of instructional decision-making in Texas I and II. – In: D. Mewborn; P. Sztajn; D. White (Eds.), Proceedings of the Twenty-fourth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education PME-NA24. Vol.2. Athens,GA, October 26-29, 2002, p. 539-563.

Confrey, J.; Castro-Filho, J.; Wilhelm, J. (2000): Implementation research as a measure to link systemic reform and applied psychology in mathematics education. – In: Educational Psychologist 35(No.3), p. 179-191.

Confrey, J.; Makar, K. (2002): Developing secondary teachers' statistical inquiry through immersion in high-stakes accountability data. – In: D. Mewborn; P. Sztajn; D. White (Eds.), Proceedings of the Twenty-fourth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education PME-NA24. Vol.3. Athens,GA, October 26-29, 2002, p.1267-1279.

Confrey, J.; Makar, K. (in press): Critiquing and improving data use from high stakes tests: Understanding variation and distribution in relation to equity using dynamic statistics software. - In: C. Dede (Ed.), Scaling Up For Success: Lessons Learned from Technology-Based Educational Improvement. Boston, MA: Harvard Press.

Erickson, T. (2000): Data in depth: Exploring mathematics with Fathom. – Emeryville, CA: Key Curriculum Press.

Finzer, W. (2001): Fathom Dynamic Statistics (Version 1.16). - Emeryville, CA: KCP Technologies. - IBM PC. Also available for Macintosh.

Gray, J. (2000): Teachers at the center: A memoir of the early years of the National Writing Project. - Berkeley: The National Writing Project.

Heubert, J.; Hauser, R. (Eds.) (1999): High stakes: Testing for tracking, promotion, and graduation. - Washington, D.C.: National Academy Press.

LaChance, A. (1999): Promoting reform in mathematics education by building content knowledge, technological skills, and teacher community. - Ithaca, NY, Cornell University, Doctoral Diss.

LaChance, A.; Confrey, J. (2003): Interconnecting content and community: a qualitative study of secondary mathematics teachers. – In: Journal of Mathematics Teacher Education 6(No.2), p. 107-137.

Lemann, N. (2000): The big test: The secret history of the American meritocracy. - New York: Farrer Straus & Giroux.

Lieberman, A.; Wood, D. R. (2003): Inside the National Writing Project. - New York: Teachers College Press.

Makar, K. (in preparation): Developing statistical inquiry: Secondary math and science preservice teachers' immersion in analysis of high-stakes assessment data using dynamic statistical software. - University of Texas at Austin, Doctoral Diss.

McClain, K.; Cobb, P. (2001): Supporting students' ability to reason about data. – In: Educational Studies in Mathematics, Vol. 45, p. 103-129.

McNeil, L.; Valenzuela, A. (2001): The harmful impact of the TAAS system on testing in Texas: Beneath the accountability rhetoric. – In: G. Orfield; M. L. Kornhaber (Eds.), Raising standards or raising barriers? Inequality and high-stakes testing in public education. New York: Century Foundation Press, p. 127-150.

Rossman, A.; Chance, B.; Lock, R. (2001): Workshop statistics with Fathom. - Emeryville, CA: Key College Press

Authors

Confrey, Jere, Dr., Washington University in St. Louis, Box 1183, One Brookings Drive, St. Louis, MO 63130, USA. E-mail: jconfrey@wustl.edu

Makar, Katie, University of Texas at Austin, College of Education, 1912 Speedway, SZB 518F, Austin, Texas 78712, USA. E-mail: kmakar@mail.utexas.edu

Kazak, Sibel, Washington University in St. Louis, Box 1183, One Brookings Drive, St. Louis, MO 63130, USA. E-mail: skazak@wustl.edu