

Investigating Educational Practitioners' Statistical Reasoning in Analysis of Student Outcome Data

Sibel Kazak and Jere Confrey

Washington University in St. Louis, U.S.A.

Statistical reasoning is at the heart of understanding statistical concepts and ideas as well as is the foundation of empirical inquiry. It involves the use and application of statistical ideas to interpret data and make decisions based on given contexts. In other words, it is “the way people reason with statistical ideas and make sense of statistical information” (Garfield & Gal, 1999, p. 207). Along with the recent key shift in intellectual thought, permitting citizens and professionals to examine numerous complex phenomena of social importance, statistics and data analysis are becoming focal areas of mainstream school curricula in many countries. Therefore, there is a need for mathematics teachers to become more knowledgeable about reasoning with data using statistics. Further, a major shift in teachers' mathematical perspective is necessary as teaching data and statistics tends to deal with the issues of uncertainty, approximation, modeling, estimation, and prediction in relation to context, rather than to focus nearly exclusively on deduction, proof, definition and abstract mathematical systems.

Simultaneously, due to federal legislation entitled the No Child Left Behind Act, teachers and schools in the United States are being held accountable in many states based on students' mean scores or percent passing on annual statewide high stakes tests. Disaggregated results by subgroup indicate that this system of testing is having a continuing adverse impact on students, particularly economically disadvantaged and minority students (McNeil & Valenzuela, 2001). Such testing may also result in several undesirable outcomes, including the use of high stakes testing to drive school mathematics curriculum and professional development, and the acceptance of narrowly-defined content to meet more immediate, rather than long-term needs of students, thereby exacerbate the impact of disparities in educational resources.

These factors have convinced us of the importance and urgency of assisting practitioners to engage in their own investigation of data, particularly in relation to equity and to potential instructional decision-making (Confrey & Carrejo, 2002; Confrey & Makar, 2002). In 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. We conducted two research studies on statistical understanding and reasoning of educational practitioners in spring and fall 2003. We reported on common elements of both studies in Confrey, Makar, and Kazak (2004) summarizing our activities in four areas: 1) issues in the development of educational practitioners' statistical reasoning, 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 independent inquiries conducted by our educational practitioners.

The present paper will report on the fall study in which we developed and taught a one-semester course with an emphasis on assessment. In this course, educational practitioners (pre-service teachers, teachers in continuing education program, and graduate students) learned about high stakes testing and undertook studies analyzing real datasets using a statistical software tool called Fathom Dynamic Statistics (Finzer, 2001). 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 ten weeks of the course. The last three weeks were devoted to group-designed data investigations. The development of statistical understanding and reasoning of the educational practitioners and their independent investigation of data on student performances in relation to issues of equity form the central focus of this paper.

Development of Statistical Reasoning in the Instructional Sequence

The statistics instruction was organized around three key conceptual areas in which educational practitioners were encouraged to develop deeper understanding of statistical concepts and statistical reasoning in the contexts relevant to them: 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 will briefly discuss each area.

Distributions of Scores. As consideration of variation and sources of variation in data plays a central role in statistical thinking (Wild & Pfunnkuch, 1999), the first unit in statistics instruction initially emphasized understanding the idea of variation in scores and moved to systematic comparison of outcomes in the context of high stakes assessment. To motivate these ideas visually, we began by asking practitioners to compare the scores of student populations from two schools, one large and one small, on the same test and decide which student population 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 rich discussion of distributions. In particular, we found two emerging and competing approaches: 1) to partition the groups into equal numbers and compare the resulting intervals, which leads towards box plots, 2) to partition the groups into equal intervals and compare the resulting numbers in the groups, which supports the use of histograms. Figure 1 below shows the original dot plots and the two competing displays.

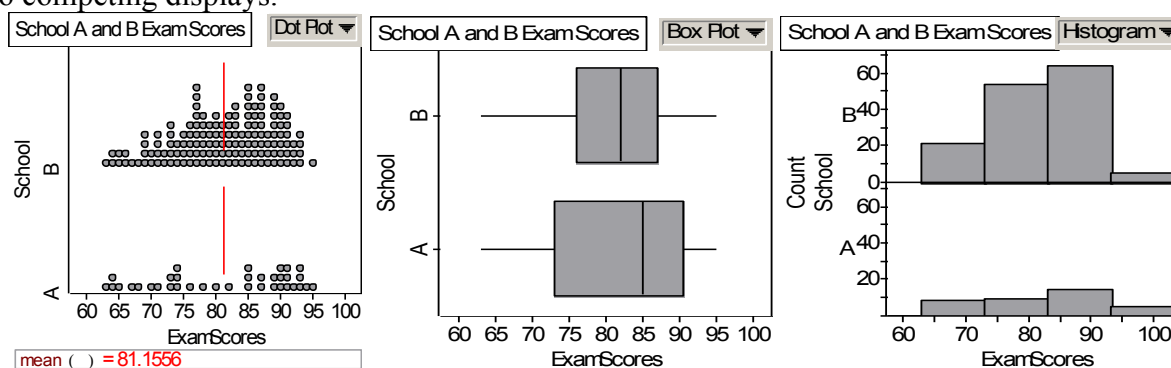


Figure 1. School A and School B distributions with the same mean and different sample sizes.

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

Next, we discussed the concept of standard deviation as a statistical measure of variability in distributions, using histograms and means. In particular, 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)-the average of the absolute deviations of values from the overall mean, which is rather an intuitive way to think of measuring variability in the data. Further, we linked the standard deviation to the inflection points in a normal distribution as a defining characteristic of that curve by transforming histograms into density curves. We discussed how changing the vertical axis to a percentage rather than a frequency does not alter the histogram's shape but does produce a display in which the total area of the bins 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 set up the transition to a distribution interpreted as a probability of outcomes.

We then assigned our students to make a normal distribution on a transparency of graph paper using 100 squares (Dienes unit blocks) and then to repeat the exercise with a different normal curve—one tall and thin, and the other short and wide. For each, we asked them to trace the shape on the graph paper and to mark the points of inflection (where the normal curve changes from falling ever more steeply to falling ever less steeply) and count the number of squares inside the vertical lines that would be created by using the two inflection points as bounds. Students were also asked to report the following distinguishing characteristics of their approximations of normal distributions: a) height of the tallest point, b) distance from the ends to the vertical center line passing through the tallest point, c) location of points of inflection relative to the vertical center line (i.e. the mean), and d) number of squares within the vertical bounds of inflection points. This last characteristic led towards a discussion of the percent of the data within one standard deviation of the mean. Accordingly, students' exploration with tracing the class of normal curves revealed that 60%-78% of the squares fell within the vertical bounds of inflection points. See Figure 2 for a simulated version of our students' explorations. After this investigation, they became convinced that if normal curves have points of inflection one standard deviation above and below the mean, then approximately 2/3 (68%) of the squares lie within this area. We then discussed how someone might have produced a formula for such a curve and argued that the general form of e^{-x^2} would produce a possible candidate. Next, we discussed how one could transform the equation to locate one standard deviation at the point of inflection and still keep the area at a total probability of one,

using the normal probability density function, $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. This was an effective way to

link the standard deviation to the normal curve and to argue for how to interpret one, two and three standard deviations in relation to the idea of a probability distribution.

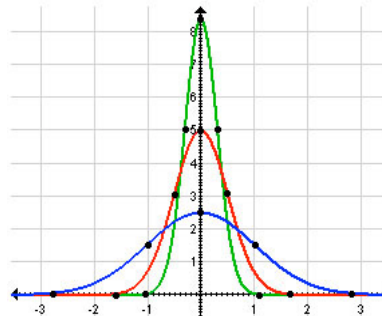


Figure 2. Investigating the class of normal curves.

In our pre/posttests, 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 test 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. The overall distribution of student responses is displayed in Figure 3, and can be summarized as follows: 1) In the pretest, of twelve students, four 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 pretest; 3) None utilized the box plots in order to compare the quartiles at the beginning of the semester; 4) Even though several students continued comparing the two distributions with the measures of center in the posttest, about 75% of them also compared the variability between the distributions; 5) Moreover, after the course, many students (58%) 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 posttest were more complete,

and the emphasis was on measures additional to central tendency or percent passing. Finally, final projects indicated that our students had developed a keener appreciation of variation and distribution of scores.

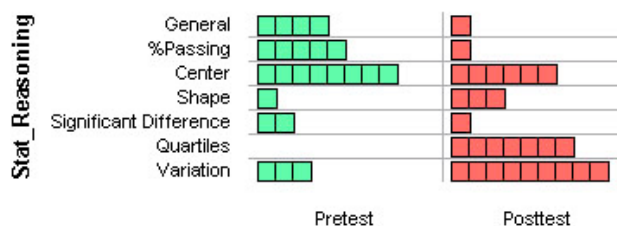


Figure 3. Student responses to comparing two box plots.

Covariation, Correlation, and Linear Regression. The second statistical unit centered around the issues of covariation—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 in the context of testing. Various representations and simulations were utilized to discuss 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.

On the posttest 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 scatterplot along with the linear regression line on the graph, the equation of the linear regression with negative slope, and r^2 . Of all students, 69% estimated the correlation coefficient by calculating the square root of the given r^2 value and taking the negative relationship between the variables into account. The rest, however, simply tried to guess from the scatterplot looking at the direction and the strength of the relationship. We found that one of these guesses mistakenly violated $-1 \leq r \leq 1$ 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.

Statistical Inference. The final statistical reasoning focus was to discuss sampling distributions and confidence intervals, and use these to develop the idea of inferential statistics. We began by exploring the notion of sampling distributions using simulations in Fathom. 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 of measures and plotted the distributions of proportion of cases that voted “Yes” (Figure 4, with “true” probability 0.5). 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 an unknown population mean on a math test using various sample sizes.

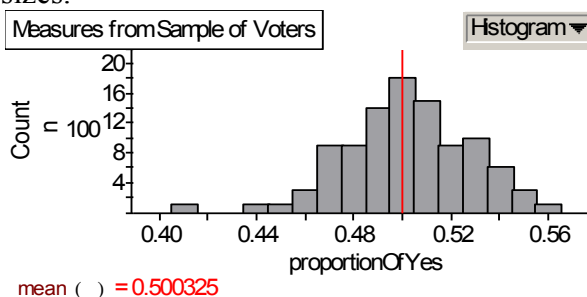


Figure 4. Sampling distribution in Fathom displaying proportion of “Yes” votes in each sample.

Two key ideas relevant to the sampling distributions emerged from student investigations: the impact of taking a larger sample of a population and drawing more repeated samples of measures 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 repeated samples of measures makes the shape of the sampling distribution more smooth and normal. It is 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.

It was relatively straightforward to move to confidence intervals as a procedure for estimating unknown population mean and developing the idea of inferential statistics because of the earlier probability-based discussions on normal distributions and sampling distributions. Specifically, our discussion of confidence intervals followed this chain of reasoning: Based on the investigation of 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 deviation 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 an examination of a sampling distribution using the difference of the mean scores seemed relatively straightforward to our students. They used it repeatedly in their projects and through repetition their use of it became secure, although we do not know about their conceptual depth. We also encouraged our students to verbalize the meaning of p-value produced by software as a way to assess the strength of evidence, i.e. “whether the sample outcome is surprising” (Cobb & Moore, 1997). In the posttest, when asked to write a conclusion statement about the given t-test output in the context of 10th grade students’ science test scores by gender, 85% of our students correctly responded to the question. Many of these responses also stated the meaning of the p-value (i.e. the observed difference is so large that it would occur just by chance only about 0.67% of the time). 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.

Types of Inquires Conducted by Educational Practitioners

Students’ independent inquiries into an issue of equity through an investigation of high stakes assessment data took place at the end of the 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. Given student-level data from nine schools in the same district in a Midwest urban area on the state-mandated test in years from 1999 to 2003, there were four groups interested in: 1) investigating how racial/ethnic backgrounds, mobility, testing accommodation, and low socio-economic status affect special-needs students’ test scores; 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 test; 3) studying disparities in math and communication arts scores on the 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 the state accountability system with the current No Child Left Behind legislation, predicting the state’s projected level of future compliance using statewide and local data, and determining the trends in student achievement on the test by disaggregated subgroups. Thus, the variety of research foci is reflected in the range of inquiries undertaken by educational practitioners.

For instance, one student reported on how mobility affects academic achievement of 3rd and 4th grade special-needs students, among whom minorities are over-represented. This inquiry was motivated by high mobility among students, especially African American (AA) and economically disadvantaged students, and the educational reform efforts aimed at improving students' academic achievements, in particular special-needs students (students with Individualized Education Programs (IEP)). The strength of this inquiry was the way the context for the statistical analyses was set up through use of the literature and systematic exploration of related ideas. In the data analysis section, the visual representations were used to display the distribution of students who were identified as IEPs in different subject areas and to show the racial backgrounds of special-needs students by mobility. After finding a statistically significant difference in the mean scores of students with and without IEPs on both math and communication arts tests by employing t-tests at a significance level of 0.001, the effect of mobility on the test scores of students with and without IEPs was investigated. In doing so, individual t-tests were performed for each pair, such as the difference between mobile and non-mobile students with IEPs and the difference between mobile and non-mobile students without IEPs. In this kind of inquiry, using ANOVA (which was not covered in the course) could be a better choice in order to see the interactions since there are two independent variables with two levels (i.e. mobile/non-mobile and with/without IEP). Analysis of mean scores of students with IEPs by the mobility status suggested that mobile students with IEPs performed better than their non-mobile counterparts. Individual t-tests, however, showed no statistically significant difference between the mean scores of mobile and non-mobile students with IEPs in either content area on the test. Since neither the standard deviations in samples nor the p-values obtained in the t-tests were reported, one must be cautious about making such a conclusion. The limitation of this inquiry, however, was due in part to the data source provided. For instance, mobility in the data sets was defined as whether or not a student was in the district and in the school for less than a year, however the data only listed "Yes" to indicate mobility and thus blank cells were simplistically interpreted as "No". Moreover, the sample sizes were fairly low when the data were disaggregated for the purpose of the study.

Another student carefully examined the variation in science achievement 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 scores among schools ($p < .05$). 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, % females, % males, % Limited English Proficiency, 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. 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 student to teacher ratio, teacher experience, and reading proficiency, which 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 ($p < .10$ 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.

Discussion and Conclusions

In the pre/posttest analysis, we looked at students' performances on four statistics items relevant to the following topics: 1) interpreting box plots to compare two groups, 2) interpreting the variability in distributions, 3) understanding the measures of central tendency and of variability, and 4) comparing the means of two distributions. Our analyses showed overall gains (Figure 5). The box plots for the distributions of scores on pre- and posttests revealed that the majority of the students performed better at the end of the course with less variation in the middle 50% of the scores. 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. Particularly, most of the gains in scores were accounted for the item on the resistance of the measures of center and variability to outliers.

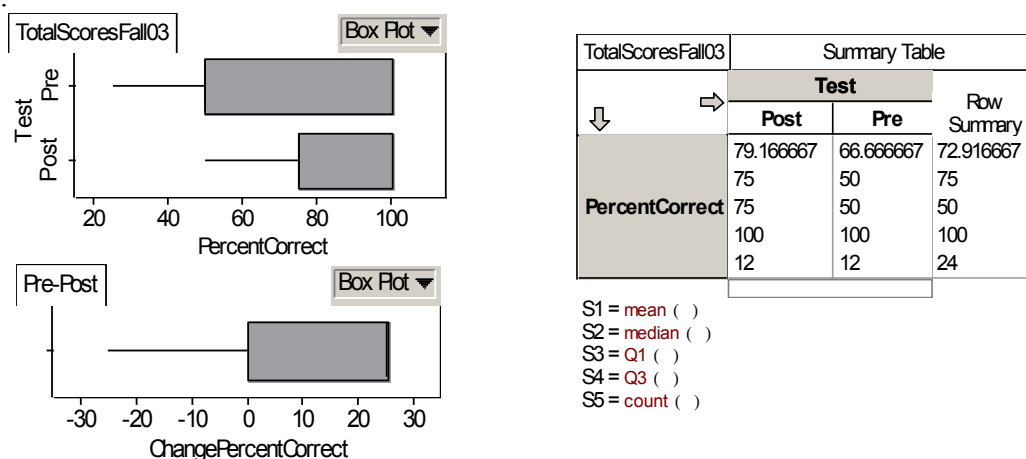


Figure 5. Distributions of percent correct in pre- and posttests with summary statistics and change in percent correct over the fall course.

Our experience with the course reveals the value of involving practitioners directly in the examination and analysis of data. Moreover, the context of data relevant to 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. Similarly, the experience in data analysis provided them with ways to strengthen chains of reasoning on issues that were otherwise sensitive to discuss. Their compelling and competent choices of investigations show that this audience was able to examine raw data and to conduct independent inquiries.

References

- Cobb, G. W. & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, 104, 801-823.
- 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 Vol.2* (pp. 539-563). Athens,GA: ERIC Clearinghouse for Science, Mathematics, and Environmental Education.
- 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 Vol.3* (pp. 1267-

- 1279). Athens,GA: ERIC Clearinghouse for Science, Mathematics, and Environmental Education.
- Confrey, J., Makar, K. & Kazak, S. (2004). Undertaking data analysis of student outcomes as professional development for teachers. *Zentralblatt für Didaktik der Mathematik*, 36, 32-40.
- 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.
- Garfield, J. & Gal, I. (1999). Teaching and assessing statistical reasoning. In L. V. Stiff & F. R. Curcio (Eds.), *Developing mathematical reasoning grades K-12 1999 yearbook* (pp. 207-219). Reston, VA: National Council of Teachers of Mathematics.
- McClain, K. & Cobb, P. (2001). Supporting students' ability to reason about data. *Educational Studies in Mathematics*, 45, 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* (pp. 127-150). New York: Century Foundation Press.
- Rossmann, A., Chance, B. & Lock, R. (2001). *Workshop Statistics with Fathom*. Emeryville, CA: Key College Press.
- Wild, C. J. & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, 67, 223-265.