SCALING UP SUCCESS

Lessons Learned from Technology-Based Educational Improvement

Chris Dede

James P. Honan

Laurence C. Peters

Editors

o

Foreword by

Ellen Condliffe Lagemann

THIS BOOK IS THE RESULT OF PAPERS COMMISSIONED AS PART OF A CONFERENCE COSPONSORED BY THE HARVARD GRADUATE SCHOOL OF EDUCATION AND THE MID-ATLANTIC REGIONAL TECHNOLOGY IN EDUCATION CONSORTIUM LOCATED AT TEMPLE UNIVERSITY'S CENTER FOR RESEARCH IN HUMAN DEVELOPMENT AND EDUCATION

JOSSEY-BASS
A Wiley Imprint
www.josseybass.com

10

CRITIQUING AND IMPROVING THE USE OF DATA FROM HIGH-STAKES TESTS WITH THE AID OF DYNAMIC STATISTICS SOFTWARE

Jere Confrey, Katie M. Makar

Data analysis can lay the groundwork for improved portrayal and interpretation of student performance.

ACROSS THE COUNTRY, proponents extol the potential use of data from high-stakes tests to improve instruction (Eisenhower National Clearinghouse, 2003). Drawing from models of business, the argument suggests that by a careful examination of overall data reports and the disaggregation of data by subgroup (race/ethnicity, socioeconomic status, gender, language group, and educational status) practitioners can monitor the overall progress in the system and devise appropriate strategies for instructional improvement tailored to local settings. In this chapter, we use a case study and subsequent work to illustrate how these approaches to disaggregation may shed some light on the relevant issues of performance differences but lack the robust application of statistical concepts and relevant theories of equity and assessment to ensure that data use is valid and fair. We outline approaches that would support scaling up through the use of more accurate and fair ways of examining distribution and difference that are still easily accessible to policymakers, teachers, and community members. Moreover, we demonstrate the value of providing teachers with

professional development experiences in conducting similar inquiries of their own, using real data and dynamic statistics software.

In Texas, schools are provided with data on the overall performance of their students scaled to the Texas Learning Index (TLI), a scaled score that is linked to the process of test equating from year to year. These scores can be compared to districtwide performance or state performance. Groups like Just for the Kids (now the National Center for Educational Accountability, www.nc4ea.org) provide schools with profiles of the highest-achieving comparison schools in the state that share the same demographics. Their summaries report on total student populations as well as totals for individual students who have been enrolled in the school for three or more years. Schools also receive these same data for a number of subgroups, including ethnic groups (African American, Hispanic, white, and Asian), economically disadvantaged students, students with limited English proficiency, and special education students. All data are reported as percentages passing. We wish to demonstrate that separating the data into groups and reporting the percentage passing on scaled scores present an inadequate representation of the situation. Using the statistical concepts of distribution, sampling variation, significant differences, and variability of performance by objective, we use data representations to demonstrate weaknesses in schools' current system of data use.

This work is particularly important when one considers that the approach to school reform in the No Child Left Behind Act is consistent with the design of the accountability system in Texas, and its regulations are modeled closely after the Texas system. Our argument is part of a larger exploration under development in a book that Confrey is writing called *Systemic Crossfire,* in which she argues that the national model of linking standards and accountability systems into a "bookends" approach to reform masks an underlying controversy about issues of content and of equity. These conflicts produce incoherence and conflict at the classroom level, and because the model is relatively silent on practice in the instructional core (teaching, learning, and formative assessment), it leaves the public blaming the failure of the system on teachers and students, rather than on the naïveté, incompleteness, and underfinanced aspects of the model, including poor funding of professional development. In this chapter, our investigations focus on how the use of feedback in that model of systemic reform suffers from inadequate approaches to the employment of data. Particularly problematic is the role of data analysis in ensuring that the system is properly attending to issues of fairness and consequential validity (Messick, 1995).

Nancy Love (2003) has undertaken some of the most well regarded studies of the use of data. She advises teachers and administrators to learn what they can from standardized tests and warns that the summary data reported to the public does not provide teachers with the information they need to guide instruction. She recommends digging further into disaggregated data to search for potential gaps between genders, socioeconomic levels, special education designations, and ethnic or racial groups—gaps in performance that are hidden by summary statistics. While we agree that examination of data needs to go beyond aggregate data, we have found that schools draw hasty and erroneous conclusions based on even disaggregated summary statistics. In this chapter, we outline and demonstrate why current methodologies for the distribution and analysis of data are insufficient from the perspective of fairness and thus lack the validity to lead to the careful design of revised instructional practices.

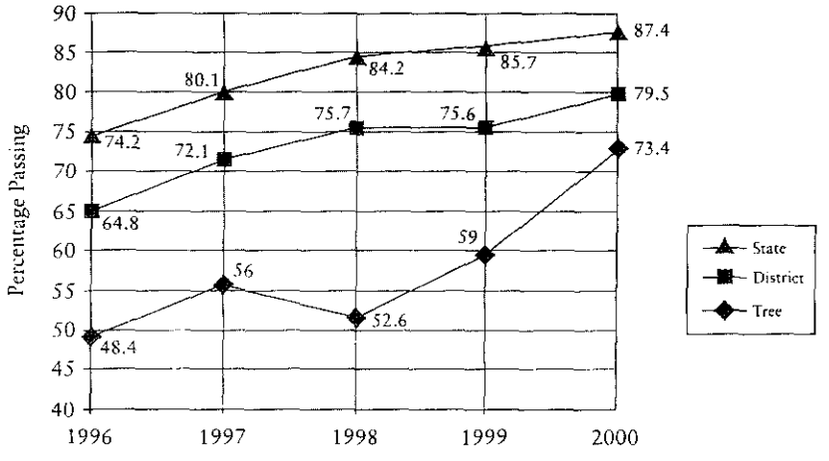
Background

For five years (1996–2000), our research team partnered with Tree High School (not the school's real name), a high-poverty urban school in Austin, with 71 percent Hispanic, 11 percent African American, 16 percent Caucasian, and 2 percent Asian and others, as well as 43 percent economically disadvantaged students. At the end of a five-year period of working with the entire mathematics department on the use of replacement units (short curricular units used by a set of teachers to experiment with or build a transition to a new approach) in algebra, our research collaborative had seen a 25 percent increase in the passing rate on the state exit test, called Texas Assessment of Academic Skills (TAAS), in mathematics. Students are required to pass TAAS in order to graduate, and schools are held accountable for their students' performance. Overall scores (for about 300 students tested per year) had increased from 48.4 percent passing to 73.4 percent over our five years of working with the school. This was in comparison with a 13 percent gain at the state level and 15 percent district gains (see Figure 10.1).

The results of the 2000 TAAS revealed that the performance of the school's African American students had dropped below acceptable levels in mathematics. The school data were reported to staff, parents, and community members in a chart (Table 10.1), which presented the percentage of students passing TAAS, disaggregated by subgroup.

Since less than half of the African American students (15 out of 31) had passed the test, based on the state accountability system, the school was labeled as low-performing. A number of obligatory steps followed. The

Figure 10.1. Longitudinal Pass Rates on TAAS Math Test for Tenth Graders at Tree High School, Austin Independent School District, and the State of Texas, 1996–2000



Data source: *Texas Education Agency, 2000a.*

teachers were called to a one-day meeting with district administrators and curriculum supervisors at which the school data were compared to the district and state data in tables that summarized the students' performance. The motto for the day was "The clearer the focus, the greater the achievement." The district administrators, including the area superintendent, the special projects director, and the mathematics supervisor, provided state, district, and school data summaries, raw data by objective on TAAS and the new algebra end-of-course exam, and an item analysis of TAAS. Teachers were promised additional resources and training to address the weaknesses in their students' performance. The meeting included virtually no discussion of the 49 Hispanic or the 6 white students who had also failed the exam.

The teachers identified what they saw as the school's problems, which, ranked from highest to lowest, included a lack of high expectations, poor attendance, lack of formative assessment, the need for challenging math problems, the need for smaller class size, language barriers, disruptive student behavior, gaps in students' competencies in mathematics, and a lack of diverse cultural representations and role models. The district ruled out the possibility of changes to class size and attendance patterns and focused

Table 10.1. *Percentage of Students Passing the TAAS Math Test, 1999-2000*

	State	District	Campus Group	Campus	African American	Hispanic	White	Native American	Asian and Pacific Islander	Male	Female	Economically Disadvantaged	Special Education
2000	86.8	81.7	82.3	73.4	48.4	74.5	85.7	—	83.3	80.4	65.6	71.3	42.9
1999	81.6	73.9	75.5	59.0	58.1	55.7	73.9	*	80.0	54.9	63.5	58.3	40.0

* *Sample size was too small for the State to report results*

Data source: *Texas Education Agency, 2000b.*

the teachers on two questions: (1) How are you using data to focus instruction? and (2) How are you differentiating your program for students who are members of the underperforming group? At that meeting, no attention was given to the distribution of scores of the African Americans in order to explore the depth of the problem, nor did any examination of the performance of the students as a whole or by subgroup over time take place. Later, the research team met with the district administrators and pointed out that the teachers' substantial progress over the past few years, as reflected in the scores of all students, had not been acknowledged adequately. Furthermore, the analysis of the performance of African American students had not been placed in time longitudinally, nor had the district examined the distribution of scores.

After the meeting, the discussions continued at the campus, culminating in a campus improvement plan (CIP). The campus plan included the following action plans regarding mathematics: "Need a systematic concentrated effort to teach/reinforce TAAS objectives. . . . Need to address specific learning needs of special populations—how populations are defined, legal requirements, and programs/resources available for each group" (Tree High School's *2000–2001 Campus Improvement Plan*, p. 6).

One priority stated in the plan was "To increase the student passing rate of the African American TAAS subgroup to at least 60%" (Tree High School's *2000–2001 Campus Improvement Plan*, p. 39). The key strategies to obtain this goal were identified: "To provide the faculty with staff development training targeting the subgroup found in the 'condition of performance,'" "to identify all African-American students in the 9th and 10th grade and assign them a peer tutor," and to "identify and address TAAS objectives which were below 50%" (Tree High School's *2000–2001 Campus Improvement Plan*, p. 39). Also included in the CIP was a stated plan to meet with teachers from a neighboring high school in the wealthier part of town, who had eliminated low performance by requiring that all African American students attend mandatory lunchtime tutoring programs.

In response, the research team wrote a letter to the new principal and the department, which stated,

> Our concern [is] that the CIP, as currently written, unfairly targets African-American students for specific interventions when such interventions would more appropriately be directed at all low-performing students at [Tree] High School. Although TAAS results are reported by racial/ethnic group for purposes of the Texas accountability system and schools can be labeled as low performing based on the TAAS scores of one subgroup on one subtest, we believe it is important that

> resources and services be targeted primarily based on student performance, not race. If racism against African Americans is a schoolwide problem, we agree it merits schoolwide attention with regard to the whole target group. However, initiatives like assigning mentors or providing pullout treatments should not be done for an entire group. Finer distinctions would be more helpful. [letter from J. Confrey to Tree High School principal, October 13, 2000]

In her response to this letter (October 30, 2000, p. 1), the principal agreed to change the wording in the CIP to "low-performing students" but added, "we will still focus on our target population in order to adequately address the anticipated questions from the Texas Education Agency and in order to ensure all of our students' needs are being met."

During this time period, the research team and its partners watched in frustration as the new principal and her collaborators at the district level, curriculum coordinators, and district special programs directors dismantled the partnership with the university research team. The research team engaged in an effort to interview students to document their responses to the programs; in so doing, evidence surfaced showing that at least one African American student had been reassigned to special education status before the 2001 TAAS test by counselors, without his or his parents' knowledge, and was thereby exempted from testing requirements. Our interviews increased the tensions between the research team and the school personnel. At a tense meeting in the spring of 2001, the research team asked the teachers if they wished to continue the collaboration. When the teachers decided to discontinue the projects, the research team expressed a combination of relief and disappointment at the ending of the partnership, agreeing that our involvement in the school's activities was no longer productive.

This experience left our research team puzzled and disturbed. We began to consider critically what had happened and sought to understand how, in a system that professes to support standards-based instruction, such a result could lead to the demise of a program that was showing steady improvement and success in relation to the high-stakes tests and was consistent with the recommendations of the National Council of Teachers of Mathematics standards movement. We helplessly watched as the program was replaced by a program of test preparation and pull-out programs for particular students, primarily students of color. In subsequent years, the school's performance continued to rise, with 77 percent of all students and 71 percent of African American students at the school passing in 2002; however, the number of African American students taking the test

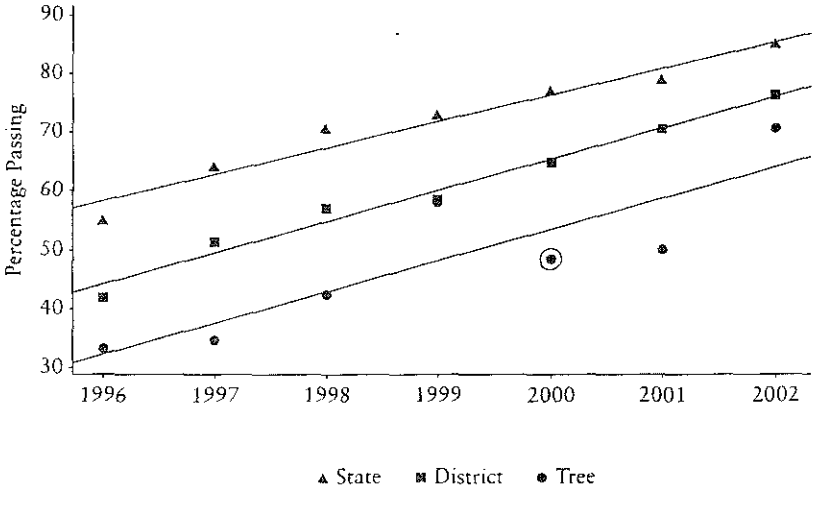
had dropped below accountability requirements in both 2001 and 2002. State and district performance also continued to rise.

Based on our observations, the school district had made a number of questionable decisions in implementing the accountability system. First, they did not look at the data in a systematic way to examine how this data point fit into the long-term trajectory, nor did they examine the distribution of the subgroup's scores and compare its performance to that of the larger population of students. Second, their treatment program was heavily oriented toward treating the students in this group as a problem to be fixed, giving little attention to other factors in the systematic treatment of African Americans in the school that might need examination. Third, the approach of "fixing" the students revealed a view of remediation that was poorly connected with a larger vision of curricular progress; it simply focused on bringing these students across the threshold of a passing score. And finally, it appeared likely that the school had adopted a strategy of ensuring that it was not designated as low-performing in the following year. Tactics included assigning students to special education or holding them back a grade, to keep small the numbers of students in certain subgroups in order to keep the school from being held accountable for their performance. This case led our team 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. It made us aware of schools' neglect of distribution, sampling variation, and inferences of statistical difference in the use of data from TAAS.

The project director presented evidence to the district administration that decisions had been based on an incomplete assessment of indicators. However, the administrators would not consider the longitudinal trajectory of the whole tenth grade or the fact that the pass rate for that year was the second highest of the past five years. Had the administration pursued a more robust study of the statistics behind the data, they would have seen a different story. First, the long-term trajectory of student performance had continued to rise during the partnership (Figure 10.1). Second, the African American subgroup continued to be on the same trajectory as the district and state when examined longitudinally, based on a least-squares fit of the data (Figure 10.2). Third, the drop in performance of the African American subgroup was not unusually large, given the high variability of past performance of the subgroup. We found that the residual of the 2000 drop in performance was still within one standard deviation of their projected pass rate of 54 percent (based on the least-squares regression line). Given the small size of the subgroup, our simulations indicate that if we accept the trajectory of performance of the

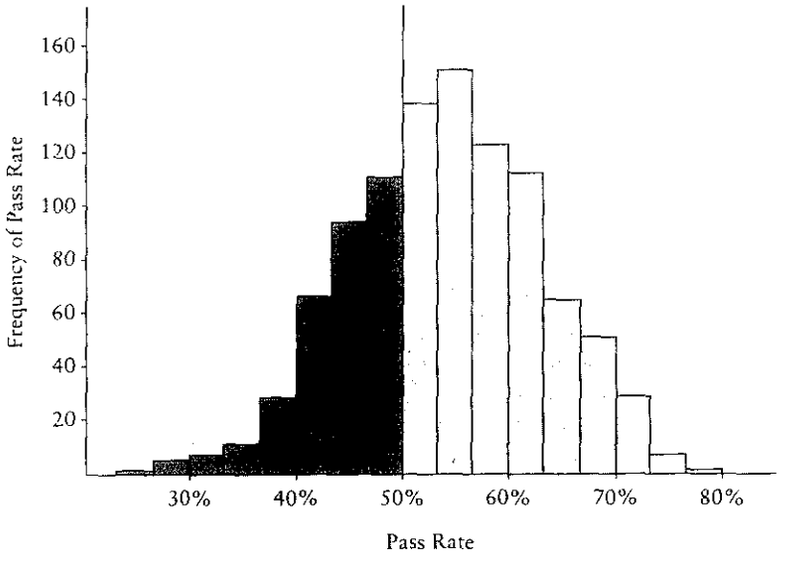
African American population over the seven years since the partnership began in 1996, the subgroup had more than a 30 percent chance of falling below 50 percent passing just by chance (Figure 10.3). Even if the district had based their decisions only on the 1996–1999 data available at the time, the 2000 drop in scores was still not statistically significant. Furthermore, in an examination of the distribution of scores of the African Americans, we found that a number of students were close to passing; in fact, if one student had answered one or two more questions correctly, the school would not have been labeled as low-performing. We demonstrate the probability of the low-performing score by using a simulation rather than a *t* test because simulation is easily supported by the software we used and because it produces a visual display that gives a spread of outcomes from a diverse set of samples and makes the point more vigorously that variation in sampling distributions is expected. Finally, in our attempt to verify the scores of the 31 African American students at the school as reported by the state, we found that only 20 were still enrolled.

Figure 10.2. Longitudinal Performance Trajectories of African American Students on the TAAS Math Test in the State of Texas, Austin Independent School District, and Tree High School, 1996–2002



Note: *The 2000 drop in performance of African Americans at Tree High School is circled. The 1999 performance of the African Americans at Tree High School overlaps with that of the district.*

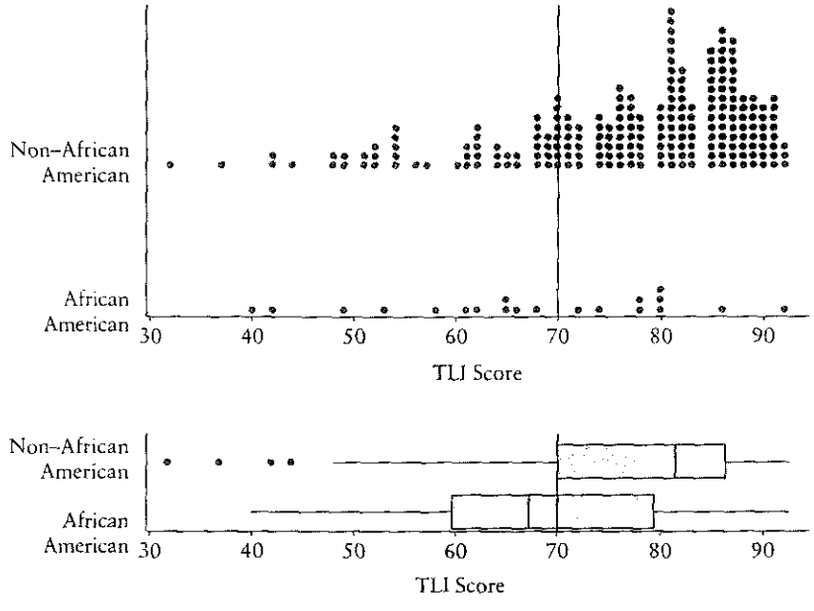
Figure 10.3. *Simulation* Showing Variation of Pass Rates for a Subgroup of Thirty-One Students When Projected Pass Rate Is on Trajectory (53.6 percent)



Note: *Due to sampling variation, in more than 30 percent of the 1,000 cases in the simulation, the subgroup of 31 students had a pass rate below 50 percent.*

In order to see a more accurate and complete picture of the performance of the African American subgroup at Tree High School, we present first a figure showing the African American students' distribution in relation to the distribution of the whole school (Figure 10.4). We represent the same data in the form of box plots as well, to assist the reader in considering how different representations support different pictures of student performance. For example, the box plot allows easier comparison of unequally sized groups, particularly of the variation in student outcomes by performance level noted by the quartiles. The dot plot, on the other hand, does not obscure the group sizes, which are needed to determine whether differences in performance may be due to sampling variation based on the size of the groups. Because these data were not available publicly, we requested them from the area superintendent. We tried to corroborate the presence of the 31 African American students by matching identification numbers but were able to find only 20 of them.

Figure 10.4. Performance of African American Tenth Graders on the TAAS Math Test Compared with the Rest of the Student Population at Tree High School, 2000



Note: *The same data is shown as a pair of dot plots (top) and a pair of box plots (bottom). The passing standard (TLI score = 70) is marked on each plot. In the box plots, the length of the box encompasses the second and third quartiles and the vertical line shows the median.*

The conclusions we draw from these analyses are twofold. First, we believe it is essential to provide a comprehensive picture of the scores of the whole student body and the identified subgroups in terms of (1) percentage passing, (2) median or mean score, and (3) distributions marked by standard deviation or quartiles. The relationships between these sets of scores must be displayed in order to show the variation among members of subgroups and examine the significance of differences in the measures of central tendency.

Second, we emphasize that scores of students represent only an estimate of their ability to perform on a test. There is inherent sampling error in any measurement taken. In particular, small subgroups demonstrate much greater variability in performance than do larger subgroups. This suggests that placing a school in the low-performing category on the basis of a subgroup of

30 students may fail to take into account expected variations in scores. While we certainly recognize that a pass rate of only half the students is cause for concern and significant intervention efforts, we believe that the policy should be based on better-informed ideas of variation and distribution.

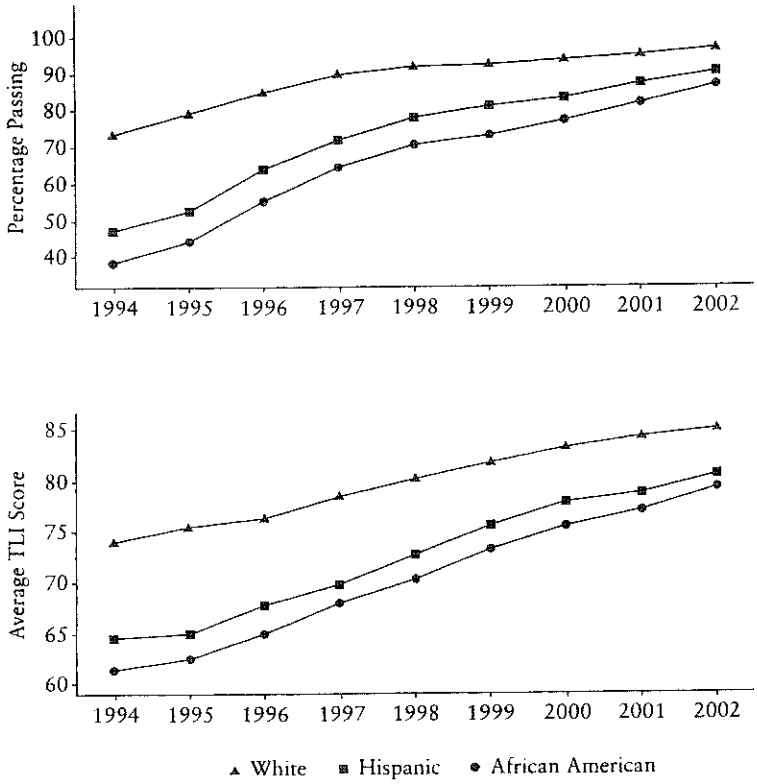
As a result of the dissolution of our partnership, we began to ask other questions about the testing system and its use of data, particularly in relation to distributions and variation. Over the past three years, we have worked with teachers to make more informed analyses of their school's assessment data (Confrey and Makar, 2002; Makar and Confrey, 2002; Makar and Confrey, 2004). In our work with teachers and data, we discovered other issues that are neglected in consideration of summary data.

State policy tends to encourage schools to focus on indicators of success rather than improvements in learning for all students. In Texas, the accountability system uses percentage passing rather than mean score as an indicator of a school's success. This leads schools to focus their energy where it will have the greatest impact on this indicator: students who are close to passing. These "bubble kids" are given additional remediation and test preparation, while other students are neglected. The statewide impact of this emphasis becomes apparent when one compares two longitudinal state-level graphs: percentage passing and average scaled score (Figure 10.5). In the percentage passing graph, the achievement gap appears to be closing as one looks at the differences in the scores of higher and lower performers over time, from left to right. However, this feature is less apparent when one examines the average TLI of subgroups over the same time period, in which the same differences narrow only slightly.

If one were to graph the individual student-level scores used to calculate the summary data in Figure 10.5 as a set of longitudinal box plots, the progress by quartiles over time would become apparent. However, the data needed to produce such a graph are not reported to the public. Analyzing this representation, one could see that the idea that forms the basis of the No Child Left Behind accountability system is faulty. Teachers and schools are not held accountable for improving the performance of the bottom portion of the scoring distribution. Attending to only the top 50 percent allows schools to avoid categorization as low-performing. Schools are also held accountable for keeping the dropout rate tolerably low, but this is not equivalent to being obliged to teach all students successfully.

Based on this analysis of distribution of scores, we argue that performance of state-designated subgroups (based on ethnicity, language proficiency, socioeconomic status, and gender) are not the only sets of student subgroups that need to be considered. For example, the lowest-testing 30 percent of students at a school are often neglected, particularly if they fall well below the passing standard. We propose that a standard be set for

Figure 10.5. Longitudinal Performance on the TAAS Math Test, Shown as Percentage Passing and Average Scaled Score, for All Tenth Graders in Texas, 1994–2002



Note: *The upper graph shows a closing of the performance gap between tenth graders of different races over time when performance is viewed in terms of the percentage of students who passed the math portion of the TAAS exit exam. However, when performance is viewed as the average TLI score on the same exam, narrowing of the gap between subgroups is less apparent.*
Data source: *Texas Education Agency, 2002b.*

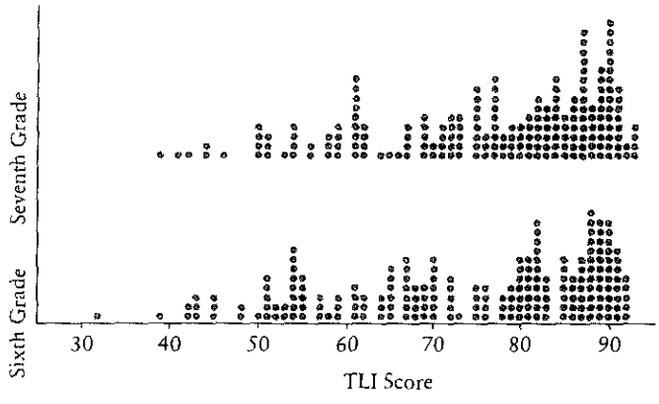
the improvement of each quartile rather than a simple report of percentage passing; if indeed there is a commitment to the adage "all students can learn," such measures should be compulsory.

In addition to focusing on the variation in scores during a single year, it is important to look at the performance of individual students over time to see the extent to which students' performances change within the distributions. For example, in the pair of graphs in Figure 10.6, the

performance of students on the TAAS exam in sixth grade are compared with the performance of the same students in seventh grade. One might assume that an individual student's performance relative to the group would remain somewhat the same from year to year—that is, students who are clustered close to passing in seventh grade would be the same ones who had scores close to passing in sixth grade.

A dynamic software program, Fathom (Finzer and Swenson, 2001), can help in such analyses. A feature of Fathom is the ability to select a subset within one distribution and show the same sample highlighted within other distributions. Using this feature, we can see that in fact, the students in the highlighted cluster in seventh grade came from a wide range of scores in sixth grade (Figure 10.6). This variation in the performance of particular students raised significant questions about what is being measured for these students each year. Conducting similar analyses across students from different percentiles may reveal that test results vary more among particular subgroups of students on the basis of test motivation. Theories to explain such differences could lead to intervention practices that go well beyond simple practice test items. Representations available through Fathom can highlight such neglected issues and can provide teachers with the resources to more completely analyze their own data and conduct their own inquiries.

Figure 10.6. TAAS Math Scores of Students from One School in Seventh Grade and Sixth Grade, 1999–2000

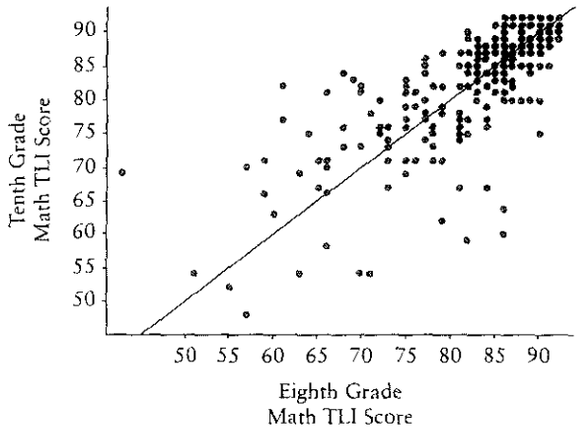


Note: *Students' scores in seventh grade are compared with their scores the year before. Fourteen students who were close to passing in seventh grade are highlighted in order to investigate their previous performance. Over one-third of them were high performers the previous year.*

Similarly, Figure 10.7 shows the performance of students on TAAS across two test administrations from the eighth to the tenth grades. The students who fall below the line $y = x$ scored lower on the tenth-grade test than on the eighth-grade test. These graphs suggest that quite a few students perform differently on TAAS across the grades. These data suggest that one would be particularly incorrect to assume that lower-performing students maintain their performance level as a group over time. Though many students lost ground over the two-year period, a significant number also performed substantially better. These two figures support a variety of interpretations, all of which, we suggest, should be considered in the professional work of school personnel serving students.

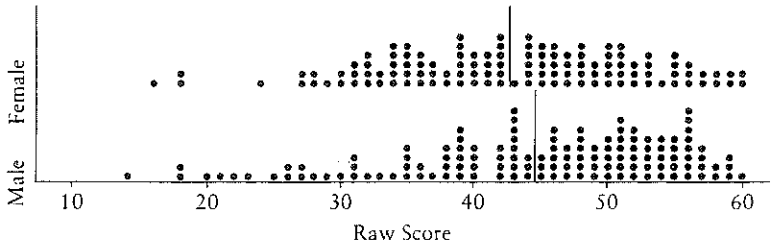
Examining data in terms of distributions raises the question of whether school personnel have been adequately prepared to draw conclusions concerning differences between groups. For example, how would they know whether a two-point difference in means between genders on the TAAS in math is a significant difference or simply due to issues of natural variability (Figure 10.8)? A permutation test can be used to compare a measure in the actual data with its likelihood of being due to random variation. The process randomly shuffles the comparison attribute (for example, gender) in the data to simulate a null hypothesis, then calculates the measure in question for the shuffled data. This is repeated multiple times, and these calculated measures are plotted as a sampling distribution

Figure 10.7. Performance in Eighth Grade and Tenth Grade of Three Hundred Students Chosen Randomly Across the State



Note: *The line $y = x$ is plotted in order to examine the pattern of change in scores. Students with scores above the line improved their scores in tenth grade, while those whose scores fall below the line did not.*

Figure 10.8. Raw Scores on TAAS Math Test from a Statewide Random Sample of Three Hundred Students, by Gender



to get a sense of their variability, so that one can compare the original measure within this sampling distribution. The Fathom software makes this process very easy to visualize. Running permutation tests multiple times, with the gender of students randomly scrambled, and noting the difference in means gives an idea of the variability one can expect. By running this type of simulation, as opposed to simply a *t* test, one can plot a graph similar to the one in Figure 10.3 (except that the horizontal access would show the difference between the means of each simulation instead of passing rate) and thus visually identify the likelihood that the two-point difference between the two groups is due to chance variation. Examining distributions without taking random variation into account or considering sample size can lead to an interpretation of a two-point difference as meaningful when it may just be due to random variation. In this case, for example, running the simulation five hundred times (similar to what was done in the simulation discussed in Figure 10.3) indicates that males performed at least two points higher than females in only about 25 percent of the cases.

To understand differences between groups, practitioners need to understand that even with statistically significant differences, there are often large amounts of overlap between the distributions of the two groups, and thus it is an act of stereotyping to assume that all members of a population perform at the lower or higher mean. Statistical significance is more easily obtained with higher sample sizes, and decisions in practice still require an interpretation of whether those differences constitute not just statistical but also educational significance. Finally, it must always be kept in mind, as we discussed in the example concerning African American students, that sampling variation can produce the appearance of difference, but this finding may not hold up under close examination of the variance of the sampling distribution.

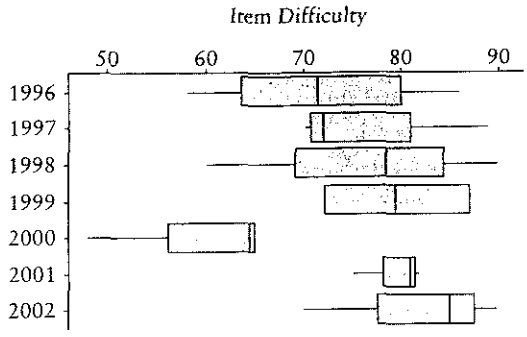
So far, our discussions of distribution and variation have concerned differences in students' total scores. Yet there is also the key question of

variation of scores on the objectives of a single test and perhaps variation in the performance of subgroups on the different objectives. Teachers who focus on teaching particular topics find this data by objective particularly compelling in helping them to make instructional decisions. In fact, at the retreat with our research group, district administrators spent a great deal of time identifying the lowest performances by objectives and stressing that the teachers needed to concentrate on teaching these topics to the low-performing subgroup. In addition, the campus improvement plan identified this as a critical area for intervention. However, objective-level data are reported as the average number of questions pegged to a particular objective that students answered correctly, based on a raw score. Because the difficulty levels of the questions vary from year to year, comparing these averages can be exceedingly misleading for teachers (see Figure 10.9). How can they know whether students' drop in performance is due to difficulty with division or due to more difficult questions presented under this objective? Conversely, there is no way for teachers to know whether performance gains are due to their increased instructional efforts or due to easier questions (Confrey and Carrejo, 2002). Figure 10.9 illustrates this problem; it shows the difficulty of questions (based on the percentage of students that answered each question correctly) under objective 9 (division) over time.

Variable difficulty of items is a key issue in data use. If the data are scaled only as a total score, then significant amounts of information that is crucial to the curriculum are lost. We argue that it is indefensible to provide teachers with unscaled data by objective if the items on those objectives are not sampled consistently from the various levels of difficulty (which could be ascertained by field testing). Not to provide teachers with valid and reliable data by objective is also unacceptable because it renders teachers unable to direct instruction and curricular treatment toward the topics for which students need additional help. This dilemma is a very serious one that is examined further in a related paper (Confrey and Carrejo, 2002) that asks questions about the validity of the psychometric construction and analysis of TAAS.

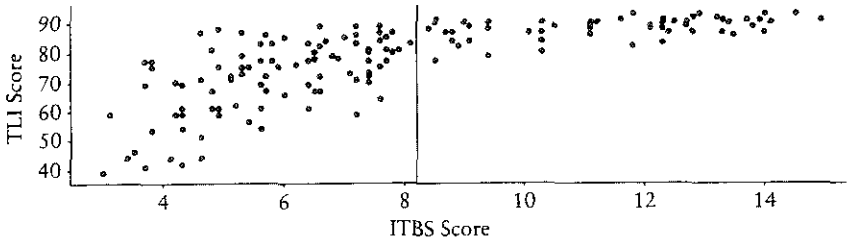
To examine the external validity of TAAS, we also conducted a brief analysis of the TAAS scores in relation to scores on the Iowa Test of Basic Skills (ITBS). The ITBS is a nationally norm-referenced, standardized test that reports student performance as estimated grade level ability. For example, an ITBS score of 8.3 indicates that the student performs at the median level of a student in the third month of grade 8. Figure 10.10 indicates a strong ceiling effect on TAAS for students performing at or above

Figure 10.9. Variation in Difficulty of the Questions on Objective 9 (Division) of the TAAS Math Test



Note: *The box plot shows the difficulty level (percentage of students answering each question correctly) of the four questions on objective 9 for each year from 1996 to 2002; it shows high variability in difficulty from year to year.*

Figure 10.10. Performance of Eighth Graders at One School on the Math Portion of TAAS and the Math Portion of the Iowa Test of Basic Skills, 2000



Note: *Grade level* = 8.2.

grade level on ITBS. However, it also indicates a weaker relationship between student performance on ITBS and TAAS for students who are below grade level. This weak relationship could indicate that one or both tests are not producing accurate readings at the lower scores or that at the lower scores, the tests are not measuring the same aspects of achievement and learning.

Discussion of Data and Results

The examples presented here demonstrate how the common forms of reporting data from high-stakes testing (for example, see Table 10.1) provide overly simplistic and potentially misleading information. Furthermore, we argue that by presenting the results on percentage passing by subgroup, without reports of central tendency and standard deviation or by quartiles, the state of Texas increases the likelihood that schools will treat low-performing subgroups as deficits to the school, contributing to racial or class stereotyping. While we recognize that the policy was intended to increase fair and equitable treatment across groups, we suggest that the failure to develop a more sophisticated view of distribution and variation is having a negative impact on racial and cross-class relations.

Interventions with little educational merit, such as those outlined in the case, are more likely to follow from this simplistic presentation of disaggregation; such interventions include the following:

- Treatment of certain conditions based on race rather than performance level
- Pull-out programs targeted at particular students that detract from their overall education
- Special efforts focused on "bubble kids" whose performance may not show consistency over time and hence may not need incremental improvement
- Overuse of practice tests
- A tendency to ignore the lowest-scoring and most needy students until high school, at which point they are denied diplomas

Since school practitioners are not held accountable for all students, the lowest-performing group over time is not required to improve. By the time this group reaches high school, expectations for remediation have become increasingly unrealistic. There is considerable debate about how many of these students drop out (Supik and Johnson, 1999) and how to calculate the number of dropouts in systems marked by high student mobility. Current predictions by the National Center for Education Statistics rank Texas twenty-sixth out of forty-seven states that reported, setting the rate of dropouts at 5 percent per year, which would translate into approximately 19 percent per high school cohort. This number has been reported to be as low as 9 percent by the Texas Education Agency and as high as 42 percent by the Intercultural Development Research Association (Supik and Johnson, 1999). According to the current accountability model, students denied diplomas due to results of high-stakes exams do not count

as dropouts. As shown in Figure 10.6, considerable variation characterizes the performance of the "bubble kids" from one year to the next, variation that may be an indicator of children who are not consistent in their engagement and identification with the system. Their scores may be a function of the degree of alienation from school that they feel, an expression of a lack of confidence, the result of low expectations, or the result of differences in instructional quality. It may represent anxiety about or ease with test taking or detachment from the technologies of school testing. Whatever the explanation is, only by systematically examining the data for variation as well as the distribution of the data, as demonstrated in the preceding examples, can we make better conjectures and build more successful, focused approaches.

Finally, it is in the failure of the system to provide valid and reliable data at the level of specific curricular objectives that we see the most disturbing aspects of a lack of examination of distribution. There is no way to use the current system to gauge students' and schools' progress at the level of particular curricular objectives. Three losses accrue from this predicament: (1) we cannot use the results by objective to improve instruction; (2) when summed across objectives, a student's score may represent a kind of overall capacity to collect superficial math skills rather than real proficiency, reflecting general school competence rather than precise mathematical reasoning; and (3) we are thereby unable to truly determine whether the system is improving. When one sees that on TAAS, the correlation between the math and reading tests is $r = 0.68$ (Texas Education Agency, 2002a), concern about the second claim is strengthened. And the refusal of the testing companies to release publicly the items they use for test equating makes it impossible to determine whether the system is improving or whether, through equating and narrowness of item sampling, there is so much drift toward predictability in the system that only improved test performance and not growth in knowledge is occurring.

The failures of the system to recognize, diagnose, and treat these problems may have produced a system in which some educational leaders can claim success and progress, but the imminent implementation of a new test, the Texas Assessment of Knowledge and Skills (TAKS), will reveal the cost of an improperly designed system of accountability.

TAKS is a more difficult exam; the math portion tests high school students' knowledge of algebra, geometry, probability and statistics, and ratio reasoning. In addition to the math test, students must pass the other three tests in science (physical science, biology, chemistry, and physics), history, and English language arts in order to graduate from high school. The projected failure rates are alarming. In the fall of 2002, the Texas Education Agency's Passing Standards Panels recommended a passing score of 33

out of 60 items, or 55 percent correct on the TAKS mathematics exit exam, which is administered to eleventh-grade students. If the State Board of Education had passed this 55-percent standard, the field test data indicated that 173,600 out of 280,000 students, or 62 percent, would have failed the mathematics test, which is required for graduation. For African Americans, the predicted failure rate was 80 percent; for Hispanic students, 73 percent; for whites, 53 percent. Economically disadvantaged students' failure rate was predicted to be 75 percent. Experience with the TAAS exam suggested that under high-stakes conditions, scores would improve by approximately 15 percent. The board relaxed the standard for the 2003–2004 year to 41.7 percent correct, or a score of 25 out of 60, for which a failure rate of 117,600 students, or 42 percent, was predicted. These results raise serious questions as to whether the previous system has moved Texas into a state of readiness for new tests. At the standard-setting meeting of the State Board of Education and the Texas Education Agency on November 15, 2002, Sandy Kress, a White House adviser in education, testified in favor of implementing the passing scores as recommended by the panels and offered the following response to a board member's question about why the scores on the field tests were so low:

> I will be candid with you. The board set the TAKS curricula five years ago, but, candidly, I don't think it's been truly taught. I think it's beginning to be taught, but I think that we can make the argument that these new standards should have been set a year or two ago, that we may be late, not early. Once our teachers, our parents, and our community understand what's expected, they'll rise to the occasion, just as they did on TAAS. So my first answer is that you probably haven't seen the implementation of TAKS. My judgment is that if you postpone setting standards, you will have the inadvertent effect of further delay before implementation of TAKS. We have to get on down the road. We have to pass standards.

Later, Kress added,

> We've gotten there together because we put pressure on ourselves. There was a lot of pressure when TAAS came out. The passing rates were worse than the ones we are talking about now, and we responded, and Texas has done better because of it. Now is not the time to lose heart. Now is not the time to stop climbing up the mountain. I don't know where the bell curve is. I don't know where the right answer is. I believe this board has said this curriculum ought to be mastered by all children.

Our question to an administration that is so publicly committed to the use of scientifically proven programs is this: Are the data in fact being used accurately, fairly, and completely to ascertain the consequential validity of test results? Our interpretation of the situation is that flaws in the construction of the previous tests (Confrey & Carrejo, 2002) and the poor quality of released summary data have contributed to the *appearance* of widespread progress across the state. However, the errors and flaws in the design of the accountability system and the failure of the Texas Education Agency to release test-equating items make the actual level of progress virtually impossible to determine. We endorse the importance of accountability, with its potential to improve schooling, but the system of accountability must itself be held accountable for its unintended outcomes. Our criticisms rest with the lack of quality use of available data and the reluctance of testing proponents to carefully examine the consequential validity of the system. Our obligation is to remedy the deficits rather than to minimize, deny, or disregard the evidence.

In addition, we are particularly concerned because of other issues that threaten the validity of the test from the perspective of those of us who are content experts (Confrey & Carrejo, 2002). Besides questions of validity, we are concerned about the tendency to assume that a test is, indeed, an accurate measure of student proficiency, when performance on tests may be influenced by other factors such as achievement motivation, opportunity to learn, and quality of instruction. Attention must be devoted to developing the means by which the issues of distribution and variation can become more accessible to practitioners, policymakers, and community members. In the final two sections of this chapter, we report on current efforts to effect such changes.

Teachers as Inquirers

Our work with data has led us to consider the importance of providing teachers with a deeper understanding of assessment, equity, and use of data to inform instructional decision making, without expecting major investments in learning complex statistics. In this chapter, we have not discussed issues surrounding the use of classroom-embedded or formative data, but these can be considered as an extension of the ideas addressed in our work with high-stakes data.

Examining the distribution of scores provides great insight into the performance of students and subgroups of students, but this insight comes only after considerable and repeated experience with looking at distributions of data. We find that teachers who are novices at performing data

analysis tend to focus on individual students and on mean scores and pass rates, ignoring the distribution of student performances (Confrey and Makar, 2002). For this reason, teachers need instruction in the concepts of variation, distribution, sampling, and difference, and they need experience in handling and interpreting student data.

In addition, we are currently extending previous work in teaching teachers to examine their own data for patterns and trends. We report on previous studies with a small group of middle school teachers who undertook investigations of their own after a set of professional development workshops (Confrey and Makar, 2002). Our design for teachers-as-investigators was modeled after the National Writing Project; based on that example, we developed the philosophy that the best way to teach mathematics teachers to include data, statistical concepts, and technology in their instruction was to provide them with an authentic experience in acting as inquirers themselves. At the end of the pilot project, teachers posed, investigated, and documented statistical evidence related to statements such as the following:

- Practice TAAS tests are not taken as seriously as the real TAAS test and are of limited value, even for the students who need them most (remedial students).
- The schools in our district remediate students at a much higher rate than those in other districts of similar size.
- Test objectives that are problematic at the exit-level (tenth-grade) TAAS are also problematic at lower grades. Furthermore, if students do well on the problem-solving objectives, they are highly likely to pass the overall test.
- For "typical" (middle 50 percent) students, there is a strong relationship between the reading and math portions of the TAAS exam.
- The relationship between student performance on the TAAS and performance on the Iowa Test of Basic Skills (ITBS) is rather weak; in addition, the passing level for TAAS is well below grade level on the ITBS. Furthermore, students who do well on ITBS will almost certainly pass TAAS, but the converse is not true.

Makar (2004) further extended this research to investigate the use of similar approaches to assessment, data, and inquiry with a set of preservice secondary-level teachers of science and mathematics. In her research, these preservice teachers undertook investigations similar to ones given

as examples in this chapter. Immersing prospective teachers in inquiry into student assessment data will provide them with a deeper understanding of students' thinking skills and help them learn more effective ways to evaluate the learning of students at various levels of performance; learn how to consider making changes in a classroom as a function of student gains over time, differences among subgroups, distributions, and variation; and become more fluent with the concepts and technologies of statistical reasoning and inquiry. Our research suggests that scaling up must include focused efforts to build professional capacity to analyze data provided by the state and to carry out investigations of student performance in local settings in relation to theories of assessment and equity.

Toward a Theory of Distributed Equity and Steady Improvement

Even careful use of data and application of concepts of distribution, variation, and sampling by themselves will not remedy the inequities in the system, nor will they prescribe appropriate actions. As we saw with the case study involving African American test performances, when data are translated into actions, racism and systematic bias may creep into the system. We see here in Texas, as well as in other places, a failure to employ adequate and robust theories or frameworks of fairness; this hampers intentions to use accountability as a lever for equity. We do not dispute the fact that disaggregation of data can be useful in shining light on pockets of extremely unfair practices or failures to educate, but we assert that current portrayals of comparative performance are too narrowly defined and too thin on analysis.

The literature on equity documents a movement to advocate for the use of multiple measures in all high-stakes decisions, in accordance with the recommendations of the National Research Council (Heubert and Hauser, 1999). Under such an approach, one set of measures can be used to compensate for performance on another measure, so that overall performance does not rely on any single measure (McNeil and Valenzuela, 2001; Valenzuela, 2002). In mathematics, there can be little doubt that practices such as timed tests, the use of multiple-choice items, or heavy reliance on language can disadvantage some students and benefit others. More specifically, in mathematics, we see significant differences among children in their likelihood of using visual representations as opposed to symbolic or numerical ones, being able to reflect verbally on their learning, or recalling long and complex strings of formulas. The use of multiple measures clearly

contributes to equitable education, because it reduces tendencies to unduly privilege certain forms of canonical knowledge in ways that are of limited or varied relevance to successful pursuit of a field of inquiry or career.

Besides the proactive arguments for multiple measures in assessment, we find that literature on equity is dominated by identification and ways to eliminate negative practices. Thus, the research on deficit thinking (Valencia, 1997) and subtractive schooling (Valenzuela, 1999) all demonstrate convincingly and profoundly how teachers' actions in schools can systematically depress children's performance by making negative assumptions about their potential to succeed or by denying them the use of their own culturally rich heritage and resources as positive contributors and facilitators of learning.

What we are proposing differs markedly and may add strength to those efforts to eliminate bias. We believe that it is essential to provide teachers with an explicit framework for engaging with diversity of performance in classrooms, particularly in mathematics, where assumptions about variations in abilities often undermine reform efforts. Teachers must be guided in how to act on trends in students' performance without reifying them into stereotypes. They must be advised on how to handle the increasing diversity in their classrooms without simply targeting instruction to a modest and possibly nonexistent middle group, which may not represent real students. Teachers must enact approaches that permit recovery and reentry by students rather than promote high levels of attrition by at-risk performers. And teachers must know how to continually challenge high performers without leaving their peers frustrated or lost. We claim that teachers must not only be versed in the fundamental concepts of distribution, variation, and difference but also receive explicit guidance in varied but mutually supportive instructional approaches that meet the needs of a diverse student population.

To explore these ideas, we created a prototype of a simulation tool called distributed equity and steady improvement (DESI). This tool created a means of gauging progress in student learning as a function of its distributed impact on students: (1) across topics, (2) across subgroups, (3) across time, and (4) across units of analysis (student, class, school or school type, district, region, or state). DESI was designed to permit inquiry on how to differentiate systematic progress from random variation or standard error. This tool enables theory building and empirical monitoring. As a theory-building tool, it permits entry of data for sets of students with different levels of prior knowledge and different speeds of learning. These differences could be assigned randomly or according to predetermined characteristics of groups of students. If further developed, DESI

could permit consideration of how different instructional treatments might affect those students in terms of knowledge and rate of learning and could project the resulting impact on the performance of students in terms of distribution and difference. As an empirical monitoring tool, it could also permit examination of the actual distributions over time and provide feedback on their implications for individual student learning in terms of distribution and difference.

A simulation tool like DESI would have the potential to do the following:

- Track students' progress in educational settings as a product of what they know and at what rate they learn new material
- Monitor changes in overall progress and by subconstruct and by subgroup over time as indicated by (1) changes in the quartile performance of subgroups and (2) changes in the distribution of scores as a whole group
- Monitor the accountability system to see how students perform over time on individual core concepts
- Provide comparisons among groups on factors such as race and gender in terms of distribution, means, and tests for significance of difference; if differences occur, track the changes in differences over time
- Provide comparisons between performance groups by quartile over time
- Track losses in population due to failure or dropout by adjusting the measures of success to reflect the losses

Once the monitoring system is constructed, instructional treatments could be interpreted in terms of their potential impact on student learning in regard to coverage and rate of learning; instructional treatments could also be examined in relation to performance by all levels of students. A trajectory of distributed progress could be projected and compared with data on student performance as predicted through simulations.

We are seeking to describe a means of appraising a kind of distributed justice in a recursive way across classrooms, schools, districts, and states to provide policymakers, practitioners, and community members with a more complete and fair portrayal of the progress of the educational system. We seek to investigate how to use the educational system as a means of meeting the needs of all children fairly over time and distributing classroom resources, including teacher time, effectively in order to ensure reasonable progress for all children. In mathematics, we must create

frameworks that support the development of talent; in this subject area, talent is fragile and needs considerable attention and nurture. Indeed, approaches to equity in the classroom must ensure that talent, not privilege or competitive advantage, is fostered. Gaining competitive advantage refers to implementing practices that propel certain subgroups of students forward while ignoring the needs of others or making them a lower priority. Our mathematics and science classrooms are loaded with competitive advantage for some children, but not for others. This concern for only some students is evidenced by how certain groups are permitted to migrate out of mathematics classes and how atypical it is for them to recover and successfully reenter college-bound course sequences.

The concepts outlined in this chapter—distribution, changes in performance over time, differences in performance across subareas of content, and differences in preparation—all illuminate issues that make the teaching of mathematics challenging, even in the absence of forces that suppress and distort performance. A theory of equity would have to provide guidance to teachers on how to manage the range of preparation at both tails of the distribution, the differences in rates of learning, and the differences in students' individual preferences. Our proposed solution to the problem is to monitor all instructional treatments in terms of their effects on the overall data set, as represented by their central measures, distributions, and measures of difference, taking into account sources of variation.

In summary, we argue for a framework for distributed equity and steady improvement (DESI) composed of a system of monitoring for overall progress and the distributions of outcomes across student groups and content objectives over time. We suggest that the model may be useful in evaluating the effectiveness of different instructional systems, especially if fundamental instructional components can be translated into their proposed impact on students' level of knowledge and rate of learning. Through the visual display characteristics of the new information technologies, a wider range of practitioners can gain insight into features of distribution, variation, and difference. We further suggest that the use of such capability will lead to revised theories on how to effectively achieve both progress and fairness.

REFERENCES

Confrey, J., & Carrejo, D. (2002). *Can high-stakes testing in Texas inform instructional decision making?* 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.

Confrey, J., & Makar, K. (2002). *Developing secondary teachers' statistical inquiry through immersion in high-stakes accountability data.* 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.

Eisenhower National Clearinghouse. (2003). Data-driven decision making [entire issue]. *ENC Focus, 10*(1).

Finzer, W., & Swenson, K. (2001). *Fathom!* (Version 1.12) [Computer Software]. Emeryville, CA: KCP Technologies.

Heubert, J., & Hauser, R. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation.* Washington, DC: National Academy Press.

Love, N. (2003). Uses and abuses of data. *ENC Focus, 10*(1), 14-17.

Makar, K. (2004). *Developing statistical inquiry: Prospective secondary mathematics and science teachers' investigations of equity and fairness through analysis of accountability data.* Unpublished doctoral dissertation, University of Texas, Austin.

Makar, K., & Confrey, J. (2002). *Comparing two groups: Examining secondary teachers' statistical thinking.* Paper presented at the sixth International Conference on Teaching Statistics (ICOTS6), Cape Town, South Africa.

Makar, K., & Confrey, J. (2004). Secondary teachers' reasoning about comparing two groups. In D. Ben-Zvi & J. Garfield (Eds.), *The challenges of developing statistical literacy, reasoning, and thinking* (pp. 353–374). Dordrecht, Netherlands: Kluwer.

McNeil, L., & Valenzuela, A. (2001). The harmful impact of the TAAS system of 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.

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.

Supik, J. D., & Johnson, R. L. (1999). *Missing: Texas youth—Dropout and attrition rates in Texas public high schools.* San Antonio, TX: Intercultural Development Research Association.

Texas Education Agency. (2000a). *Academic Excellence Indicator System: Campus, district, and state reports, 1996-2000.* Austin, TX: Author. Retrieved September 2000 from www.tea.state.tx.us

Texas Education Agency. (2000b). *Academic Excellence Indicator System: Campus report, 1999-2000.* Austin, TX: Author. Retrieved September 2004 from www.tea.state.tx.us/perfreport/aeis/2000/

Texas Education Agency. (2002a). *TAAS data set of 10,000 students in Texas.* Austin, TX: Student Assessment Division, Texas Education Agency.

Texas Education Agency. (2002b). *Texas student assessment program: Technical digests for the years 1996-2002.* Austin, TX: Author.

Valencia, R. (1997). Conceptualizing the notion of deficit thinking. In R. R. Valencia (Ed.), *The evolution of deficit thinking: Educational thought and practice* (pp. 1–12). London: Falmer Press.

Valenzuela, A. (1999). *Subtractive schooling: U.S. Mexican youth and the politics of caring.* New York: State University of New York.

Valenzuela, A. (2002). High-stakes testing and U.S.-Mexican youth in Texas: The case for multiple compensatory criteria in assessment. *Harvard Journal of Hispanic Policy, 14,* 97–116.