

CRITIQUING AND IMPROVING DATA USE FROM HIGH STAKES TESTS: UNDERSTANDING VARIATION AND DISTRIBUTION IN RELATION TO EQUITY USING DYNAMIC STATISTICS SOFTWARE

Jere Confrey
Washington University in St. Louis

Katie Makar
The University of Texas at Austin

Executive Summary: This paper reveals how current approaches to data use by schools, even with disaggregation by subgroup, can fail to recognize the importance of the statistical concepts of variation and distribution in monitoring systemic progress of all students. A case study in which a high poverty, majority Hispanic urban school dismantled its Standards-based reform program after the school was labeled low performing illustrates the claim. The low-performing status was based on the performance of small subgroup of African Americans (n=31 out of 271 tested) who had a pass rate of 48.4% on the Texas Assessment of Academic Skills mathematics exit exam. The authors document the failure of practitioners to consider those results in the context of its distribution and variation. We report protocols for data analysis that provide a more adequate and fair portrayal of student performance. We use a dynamic software program, *Fathom*TM (Finzer, 2001), in our demonstrations that can provide teachers with the resources to more completely analyze their own data and conduct their own inquiries.

Proponents of high stakes testing extol their potential use in improving instruction (ENC, 2003) and suggest that by careful examination of overall and disaggregated data, practitioners can monitor progress and devise tailored strategies for local instructional improvement. In this paper, we use a case study and subsequent work to illustrate how these approaches to disaggregation may shed some light on 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 more accurate and fair examination of distribution and difference, yet are still easily accessible to policymakers, teachers and community members. Moreover, we demonstrate the value of providing teachers professional development experiences in conducting similar inquiries of their own using real data and *Fathom*TM, a statistical learning package that permits users to investigate data visually and create linked representations without extensive theoretical training in statistics, making data analysis more accessible to school practitioners and policymakers (Finzer, 2001).

Texas schools are given data on the overall performance of their students, as well as disaggregated data for a number of subgroups including ethnic groups, economically disadvantaged students, LEP students, and special education students. All aggregate data are reported to teachers and the public as percentages passing. We demonstrate that separating the data into groups and reporting the percentage passing on scaled scores does not adequately represent the situation. Using the concepts of distribution, sampling variation, significant differences, and variability of performance by objective, we demonstrate weaknesses in the way schools currently use data.

Case Study. We have found that schools frequently draw hasty and erroneous conclusions based on even disaggregated summary statistics. Tree High School, a high poverty, mostly Hispanic, urban school with which our research team partnered for five years, is a case in point. After four years of consistent improvement by all subgroups on the Texas state assessment (TAAS), Tree was labeled low performing due to the poor performance of its African-American students. Texas requires all subgroups of 30 or more students to maintain at least a 50% passing rate on TAAS in order to avoid low performing status. At Tree, only 48.4% or 15 of the 31 African-American tenth graders passed the exit-level test. The district and administrative personnel called a meeting with Tree teachers and presented a Campus Improvement Plan. No

attention was given to the distribution of scores of the African Americans to see how deep the problem was, or to overall or subgroup performance over time.

District and local personnel chose to dismantle the research partnership. In making this decision, they did not look at the data in any systematic way to examine how this data point fit into the long-term trend, nor did they examine the distribution of the subgroup's scores and compare it to that of the larger population of students. Their program treated students in this group as a problem to be fixed, with little attention to other factors in the instruction of African-Americans. The approach focused only on bringing students across the passing threshold on TAAS, and did not connect with a larger vision of their curricular progress. The school appeared to adopt a more strategic approach to avoid low-performing designation the following year, retaining students or assigning them to special education to avoid having enough students to be held accountable for subgroup performance.

A more robust understanding of the data would have given the administrators a different perspective. For example, a closer examination of the data reveals: (1) the African-American passing rate for that year was the second highest of the last 5 years; (2) the long-term trajectory of student performance by all subgroups had continued to rise during the partnership; (3) the African-American subgroup continued to be on the same trajectory as the district and state, based on a least-squares fit of the data (Figure 1); (4) the drop in performance of the African-American subgroup was not unusually large, given the variability of past performance. We found that the residual of the drop in performance in 2000 was still within one standard deviation of their projected passing rate of 54% (based on least squares regression). Our simulations indicate that if we accept the trajectory of performance of the African-American population since the partnership began, this group had a 30% chance of falling below 50% passing just by chance. (5) Finally, we found that a number of African-American students were close to passing. In fact, had one student answered one or two more questions correctly, the school would have avoided low performing status.

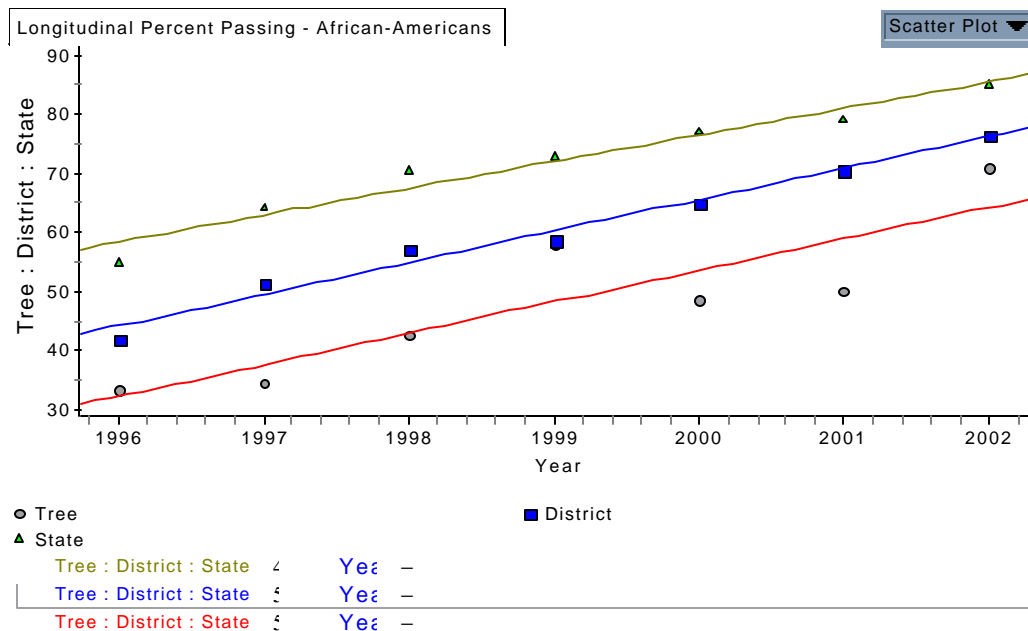


Figure 1 1996-2002 longitudinal performance trajectories of African-American students across the State (top), the district (middle), and at Tree High School (bottom). Note the 2000-drop in performance of African-Americans. The 1999 performance of African-Americans at Tree overlaps that of the District.

While we recognize that the accountability system was *intended* to increase fair and equitable treatment across groups, we suggest that lack of a sophisticated view of distribution and variation is having a negative impact on equity. This simplistic presentation of disaggregation is likely to result in:

- treatments by race of conditions that should be treated by performance level;
- pull-out programs targeted at particular students that detract from their overall education;
- special efforts for “bubble kids” whose performance may not show consistency over time and hence not need incremental improvement;
- overuse of practice tests; and
- a tendency to ignore the lowest scoring, and most needy students until high school at which point their diplomas are denied.

Using data in a more proactive approach. The paper includes other analyses of state assessment data. For example, one might assume that students at a particular performance level remain with their performance peers over time. We illustrate that this assumption is not valid: individual student scores often vary greatly over time. We show how multiple representations of data can provide greater insights into data than a single representation. We also detail an analysis of high variability of difficulty levels of items on TAAS both within topic areas over time and between topic areas on a single test, which can mislead teachers trying to gauge their instructional focus by their students’ performance on particular topics. Teachers cannot tell whether a decrease in their students’ performance in a particular year on a topic is due to lower achievement or to harder items. We also highlight the difficulty of comparing the performance of students on different types of tests or comparing subgroups on a single test without understanding variation and distribution.

These issues of distribution and variation can become more accessible if practitioners, policymakers, and community members engage in continuous monitoring of test results and discussion of the implications for students. The distribution of scores provides greater insight into performance, overall and by subgroup, but this insight comes after extensive experience looking at distributions of data. We find that teachers who are novice at data analysis tend to focus on individual students, and on mean scores and passing rates, ignoring the distribution of student performance. Thus, teachers need instruction in variation, distribution, sampling and difference and experience handling and interpreting student data. Confrey and Makar (2002) report on studies with teachers who undertook investigations of their own after professional development designed to immerse them in authentic inquiry.

To explore these ideas, we created a prototype of a simulation tool called Distributed Equity and Steady Improvement (DESI), 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. It was designed to permit inquiry on how to differentiate systematic progress from random variation or standard error. It allows one to enter a set of students with different levels of prior knowledge and different learning rates. If further developed, it could permit one to consider how different instructional treatments may affect students in terms of knowledge and learning rate and project the impact on student performance in terms of distribution and difference. A trajectory of distributed progress can be projected and compared to data on student performance from the model. It could also permit one to examine actual distributions over time and implications for individual students.

We argue for a system of monitoring for overall progress and the distributions of outcomes across groups and content objectives over time, the Distributed Equity and Steady Improvement (DESI). This model may be useful in evaluating the effectiveness of instruction, especially if fundamental components can be mapped to their proposed impact on the students' knowledge and learning rate. We argue that a wider range of practitioners can gain insight into features of distribution, variation and difference by using the visual display characteristics of new technologies. We further suggest that such capability will lead to revised theories on how to achieve both progress and fairness.

References

- 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, Athens, GA.
- Eisenhower National Clearinghouse (2003). Data-Driven Decision Making. *ENC Focus, 10*.
- Finzer, W. (2001). Fathom! (Version 1.12) [Computer Software]. Emeryville, CA: KCP Technologies.