LOOKING BEHIND THE "AVERAGE": HOW ARE STATES REPORTING TEST RESULTS

CSE Technical Report 312

Leigh Burstein

UCLA Center for Research on Evaluation,
Standards, and Student Testing

February, 1990

The research reported herein was conducted with partial support from the U.S. Department of Education, Office of Educational Research and Improvement, pursuant to Grant No. G0086-003. However, the opinions expressed do not necessarily reflect the position or policy of this agency and no official endorsement by this agency should be inferred.

Please address inquiries to: CSE Dissemination Office, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, California, 90024-1521

Questions about the meaning of reported achievement test results and whether the public is being misled are serious matters. Regardless of what educational professionals believe and want, virtually the entire non-educational sector (politicians, business community, parents, media) view testing as a valid and useful means of monitoring educational progress and see tests as viable tools in holding institutions and individuals directly accountable. They want to know the "truth" about how American students and schools are achieving.

The dilemma, as professional educators know, is that there is no one truth when it comes to assessing student achievement. Using the same measures to monitor progress (in the sense of trying to keep abreast of where we stand) and to hold specific educational units accountable raises the spectre of corrupting the meaning of the measures. Changing performance and changing achievement aren't synonymous, as Linn, Graue, Sanders (1990) and Shepard (1990) remind us. If we become too obsessed with measuring accurately the average performance of the students nationally, regionally, and locally, we can do a disservice to the educational improvement effort.

Apparently, the patterns of above norm performance that John Cannell (1987) reported are there although the extremity of the "Lake Wobegon" effect is perhaps overstated. Linn et al. (1990) point to the use of dated norms as a partial explanation for the high scores. Shepard (1990) identifies other practices that likely inflate reported test results. Two questions come to mind about what to do in light of this evidence. First, how can the use of commercial norm-referenced tests as measures of student achievement in high stakes testing contexts be changed to lead to more accurate public reporting of results? Second, regardless of how these tests or their use are changed, what are reasonable standards for "fully" reporting test results to "inform" educational improvement?

Changes in Test Use

The first question generates a laundry list of suggested modifications, both mild and strong. The list includes at least the following:

1. Accurately describe the norm group and tests administered in all documents and reports and properly "educate" the lay reader regarding the specific meaning of "average" being used by the state or district.

2. Fully describe all systematic exclusions of students from the tested group in the school/district/state and the likely consequences of these exclusions with regard to comparisons to the norms used.

3. Fully describe test administration and security procedures and their likely consequences with respect to comparisons to norms.

4. Develop guidelines/sanctions regarding appropriate and inappropriate test preparation and report evidence on local adherence to guidelines.

5. Report performance according to "annual user norms" to discourage the practice of comparison to dated norms.

6. Renorm tests more frequently (annually or biannually) and report performance only with respect to "new norm" data.

7. Use multiple commercial tests administered randomly throughout schools and districts each year to reduce the "benefits" of teaching to a specific test.

8. Develop multiple alternative forms and administer alternative forms randomly throughout schools and districts each year.

Documentation

Recommendations one through four above are essentially mild changes in the provision of information about what test results represent. Yet as Shepard's (1990) results indicate, there is certainly little uniformity across states and districts in reporting this information. From a cursory examination of the state reports provided for the Linn et al. (1990) analyses, none of the states reports details on all four areas (norms description, sample exclusion, administration instructions, test preparation guidelines). However, there are exemplary practices in documentation, as illustrated by an appendix to North Carolina's report which includes a glossary of terminology used and a discussion of the choice of reporting metric and what other alternatives were contemplated and rejected.

Clearly, brief, glossy, graphically presented results are more likely to attract positive response from the public and policy communities. Such priors dictate against including "messy" details in the body of annual district and state reports. Nevertheless, public documentation of practices, either through technical appendices or supplementary reports, should be routine practice as a means to improve the public policy discussion.

Frequent Norming

Recommendations five through eight require stronger changes that can be prohibitively expensive. Obviously, annual user norms are not hard to generate nor necessarily very expensive. On the other hand, annual or biannual norming, if done properly and carefully, would be unnecessarily burdensome for everyone involved.

What is troubling about either of the norming recommendations is that in many respects, both miss the mark of the purpose of annual collection and reporting of test results. These results are supposed to measure the status and progress of the school system. But either recommendation would change the metric for reporting frequently or add additional metrics to contend with. The standard for comparison itself becomes a frequently moving target devoid of anchoring. Trends in performance would be restricted to cross-sectional ones at a given point in time. Thus one would be replacing one bad signal (inflated performance due to test familiarity and dated norms) with another.

Multiple Forms

The use of multiple alternative test forms either from a single publisher or from multiple publishers would improve matters, since problems associated with "teaching to the test" would be lessened. However, the domain represented in these multiple forms may still be a lean one, and thus susceptible to corrosive (beating the test) testing practices. Moreover, it is unlikely that publishers would make the investments necessary to expand the number of forms they offer due to prohibitive cost; nor would they be likely to encourage the kinds of cooperative behavior that districts and states would need to administer tests from multiple publishers on a random basis.

Actually, one of the unfortunate consequences of Cannell's fixation on the Lake Wobegon-type results from commercial achievement tests is that it detracted attention from those states who employ alternative assessment strategies. Some states (e.g., California) annually administer many more test items targeted to specific curriculum areas and use this information to monitor performance and progress over

time. Item sampling techniques originally developed and used in the National Assessment of Educational Progress (NAEP; no student takes all items but all items are administered to random sample of students within schools and districts) yield a considerable amount of additional curricular detail. For example, California administered 486 mathematics test items at grade 8 from 1983-84 through 1985-86 covering 8 broader skill areas and 33 more specific ones with at least 12 items per area at the most specific level. Any school or district that can teach to this potentially broad a domain of content should raise student achievement as well as test performance.

This testing technology, in newer, more curricularly and instructionally relevant forms, should be dominating state and even district testing practices intended for monitoring and reporting progress. But it isn't, at least not in comprehensive assessment activities; rather, domain oriented assessment in schools seems to have been restricted to competency and proficiency testing of students rather than monitoring system performance.

For those concerned about normative data, work by Bock and Mislevy (1988) offers a means of scaling data from comprehensive assessments in ways that can anchor results to a given time point and, under certain conditions, can represent score scales with the types of content tasks achievable at a given score level. The California Assessment Program has been employing these methods since the mid 1980s and NAEP under the Educational Testing Service (ETS) has increasingly relied on variations of such reporting practices.

But simply expanding the number of items administered and using fancier analytical technology will not make test reporting rosy. Nevertheless, debating the technical shortcomings of these types of alternatives to current commercially available tests may be a better use of time than to continue introducing costly "fixes" of commercial test results for monitoring educational progress.

"Full" Contextualized Reporting

The root of Cannell's concern seems to be that somebody (various state and district officials, test publishers, or both) is intentionally deluding the public by reporting above average performance and harming children by falsely telling them (and their parents) that they are doing okay. While Cannell may be on the right track, his message is potentially as limiting as the practices he decries. Frankly, if all a state or district does is report the percent of their students above the national norm in a given year, the results are misleading, regardless of the test used and testing procedures. In district, state, national, or international testing programs, the practice of dwelling only on system-level average performance is simplistic, wasteful, counter-productive, and invariably misleading. To the degree that Cannell's obsession with state and district average performance detracts from the efforts to more comprehensively report their students' performance, his challenges perpetuate the worst, rather than the best, of large-scale assessment.

My point is that the single-minded concentration on the central tendency in states and districts is misguided regardless of who is doing it. It is important to contextualize the reporting of results within states, districts, and schools. Complaining that states and districts are using an easily misinterpreted metric for reporting when in fact the problem is how simplistically data are reported is misplaced rigor and piety.

As part of a feasibility study on using existing data collected by the states to construct education indicators for state-by-state comparisons of student performance at the national level (Burstein, Baker, Aschbacher, & Keesling, 1986), we urged that

auxiliary information about students and schools be used to contextualize the description of educational performance within states (and other educational units). Our analysis of state reports of assessment results (primarily for the years 1982-84) indicated that while a remarkable variety of interesting information (background, resources, curriculum and instructional activities at the student, school, and district levels) was being collected, there was little comparability in the collection and reporting of auxiliary information.

But in the ongoing development efforts regarding state level education indicators, the concern for contextualizing any achievement comparisons has become virtually axiomatic. The National Assessment Planning Project conducted by the Council of Chief State School Officers (CCSSO, 1988) devoted 5 of their 12 recommendations and well over half the report to advocate reporting (a) distributions of scores within states, (b) cross-sectional trends as changes in the proportions of students at specified proficiency levels, (c) subgroup reporting, (d) rankings by demographic variables, and (e) relating achievement to education variables. Likewise, 4 of the 12 recommendations from the NAEP Technical Review Panel report (Haertel, 1988) address similar concerns about moving beyond the reporting of system averages.

What kinds of information should states and districts be using to contextualize their reports of test results? In broad terms, three types of data: longitudinal trends, performance distributions (e.g., percentage scoring in each quarter) within and among schools/districts, and subgroup comparisons (e.g., by ethnicity/race, SES, gender, community type, language status, resource and curricular subgroups) and their cross-classifications (e.g., longitudinal trends in the proportion of Hispanic students in urban schools within the state score above the 25th percentile nationally) come to mind.

Taken in isolation, each of these types of information can be misleading and misused in much the same way as Cannell (1987) claims that overall state and district achievement test results have been. However, when combined, they provide a more accurate depiction of the performance of students in the nation's school systems. Moreover, publicly reporting achievement data in this more comprehensive and informative way would encourage better testing practices and public policy discussions about testing results. To use an old colloquialism, it is hard to hide one's dirty linen when it is all hanging on the clothesline.

State and district officials were not explicitly asked to provide information about trend, distribution, and subgroup reporting in the CRESST follow-up of the Cannell study. Nevertheless, the state reports obtained by Linn et al. (1990) were examined to determine whether these types of reporting practices had expanded and improved since our earlier study. The overall picture is still a mixed one. Generally, the practice of more refined reporting of assessment data has expanded somewhat with well over half the states reporting one of the three types of information emphasized here. With respect to trends, e.g., California juxtaposes trends from different subject areas on the same graph while South Carolina displays the percent of students falling within each national quarter over time for three grade levels. The latter display also illustrates attention to the distribution of performance within the state rather than total absorption with the average.

Washington's 1990 General Report (Washington Department of Education, 1990) illustrates what can and should be done in reporting performance distributions with a state. Figure 11 (see Appendix A) from that report presents performance distributions for students categorized by ethnic/minority status. this display uses box-and-whisker plots, a very compact and graphically appealing means of conveying distributional data. The body of the report provides a succinct and clear explanation of the technique.

Final Comment

I am enthusiastic about efforts to report achievement test data more comprehensively and generally unsympathetic toward Cannell's single-mindedness on the question of misleading reporting. There are already too many pressures to oversimplify matters. In the cases of Washington, Cannell (1987) managed to notice only the single number representing the state average among the myriads of displays and discussions that attempted to document how the students within the state were doing. It does a disservice to educational officials to ignore such efforts. Moreover, it undoubtedly slows down progress on more important education quality reporting to divert all attention to a particular limiting feature of the educational achievement yardsticks. We need both better yardsticks and better use of them.

References

Bock, R.D., & Mislevy, R. (1988). *Comprehensive educational assessment for the states: The duplex design* (CSE Tech. Rep. No. 262). Los Angeles: UCLA Center for the Study of Evaluation.

Burstein, L., Baker, E.L., Aschbacher, P., & Keesling, J.W. (1986). *Using state test data for national indicators of educational quality: A feasibility study* (CSE Tech. Rep. No. 259). Los Angeles: UCLA Center for the Study of Evaluation.

Cannell, J.J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all states are above the national average.* Daniels, WV: Friends for Education.

Council of Chief State School Officers. (1988). *On reporting student achievement data at the state level by the National Assessment of Educational Progress* (Recommendations from the National Assessment Planning Project). Washington, DC: Author.

Haertel, E. (Chair). (1988). *Report of the NAEP Technical Review Panel on the 1986 reading anomaly, the accuracy of NAEP trends, and issues raised by state-level comparisons.* Washington, DC: National Center for Education Statistics.

Linn, R.L., Graue, M.E., & Sanders, N.M. (1990). *Comparing state and district test results to national norms: Interpretations of scoring "above the national average"* (CSE Tech. Rep. No. 308). Los Angeles: UCLA Center for the Study of Evaluation.

Shepard, L.A. (1990). *"Inflated test score gains": Is it old norms or teaching to the test?* (CSE Tech. Rep. No. 307). Los Angeles: UCLA Center for the Study of Evaluation.

Washington Department of Education. (1990). *Washington statewide assessment general report: Grades 4, 8, and 10. Fall 1989.* Olympia: Author.

Appendix A

FIGURE 11. GRADE 4
Distributions of Ethnic/Minority Students' Scores
on MAT6 Total Reading – October, 1989