CRESST POLICY BRIEF 3

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

www.cse.ucla.edu

Spring 2001

Reporting School Quality in Standards-Based Accountability Systems

Robert L. Linn

Robert L. Linn is co-director of the National Center for Research on Evaluation, Standards, and Student Testing and Distinguished Professor of Education at the University of Colorado at Boulder. He is the former chairperson of the National Research Council's Board on Testing and Assessment.

It is commonplace to talk about education in the United States as 50 separate educational systems. The vast differences in state accountability systems exemplify that characterization, with virtually no two comparable assessment programs at this point in time. Unfortunately, this variation results in our learning less from research on state testing programs than we would hope.

Nevertheless, given policymakers' and educators' plans to expand accountability, we need to develop better designs, evaluations, and redesigns of assessment and accountability systems. This policy brief builds on the issues of school accountability and reporting discussed in recent CRESST work (Linn, 1998, 2001; Linn & Baker, 1999).

Measuring School Quality—Current Status

Increasingly, states are using their accountability systems as much for measuring school status as for student achievement. According to a survey by *Education Week* (Orlofsky & Olson, 2001), 27 states assign ratings to schools or identify low-performing schools. President Bush's education plan (2001) would require "school by school report cards…for all public schools," and publication on the Internet. But the methods by which states rank schools or measure improvement vary greatly. The most common method of reporting school results is in terms of current status, often done by reporting the school mean or median score for students in the grade assessed. States have been moving in recent years away from the use of scale scores and percentile ranks to the percentage of students who meet or exceed a performance standard or the percentage of students in each of several performance categories. The Florida school accountability system (Table 1), for example, grades schools from A to F based on current performance of students on the Florida Comprehensive Assessment Test (FCAT). Florida also reports performance in terms of improvement.

Table 1 shows that grades C through F are determined solely by student performance during the current year, whereas grades A and B have added requirements for the performance of subgroups of students, and grade A has requirements for year-to-year change as well.

Table 1. Rules for Assigning Grades to Schools in Florida

Grade	
A	Meet grade "B" criteria AND the percent of students absent more than 20 days, percent suspended and dropout rate (high schools) are below state averages AND there is substantial improvement[1] in reading AND there is no substantial decline[2] in writing and math AND at least 95% of standard curriculum[3] students were tested.
B	Current year reading, writing, and math data are at or above higher performing criteria, AND no subgroup[4] data are below minimum criteria, AND at least 90% of standard curriculum students were tested.
C	Current year reading, writing, and math data are at or above minimum criteria.
D	Current year reading, or writing, or math data are below minimum criteria.
F	Current year reading, writing, and math data are below minimum criteria.

Note. Grade description criteria and footnote quoted from Florida Department of Education (1999, pp. 1-2).

[1]Substantial improvement in reading means more than two percentage points increase in students scoring in FCAT levels 3 and above. If the school has 75% or more students scoring at or above FCAT level 3 AND not more than two percentage points decreas from the previous year then substantial improvement is waived.

[2]Substantial decline means five or more percentage points decline in the percent of students scoring FCAT achievement level 3 and above in math OR five or more percentage points decline in the percent of students scoring 3 or above *Florida Writes!*

[3]Standard curriculum students also include Language Impaired, Speech Impaired, Gifted, Hospital Homebound and LEP students who have been in an ESOL program more than two years.

[4]Under current rule subgroups include economically disadvantaged, Black, White, Hispanic, Asian, and American Indian students.

Table 2 defines Florida's minimum and "higher" performing criteria referred to in Table 1. What appears to be a fairly straightforward A-F reporting system for current school status is considerably complex in its details. Most state accountability systems are at least this complex. Like Florida, each has different features making none of them comparable.

Measuring School Quality—Improvement Over Time

A preferable approach to measuring and reporting school achievement is to place greater emphasis on improvement than on current status. A common method is to compare test scores between two years but for the same grade, for example, third-grade reading in 1998 to third-grade reading in 1999. Such "improvement-over-time" comparisons based on successive groups of students in selected grades are reasonable for schools with consistent student populations. Williams School, a fictitious name but with actual school results, shows reasonable improvement in Grade 3 reading proficiency with slightly higher "proficient or above" scores in 1999 than 1998 (Table 3). The validity of inferences from such comparisons is questionable, however, for schools with rapidly changing demographics or with too few students tested in a specific grade.

Longitudinal and Quasi-Longitudinal Reporting Methods

Another way to measure improvement is to track the performance of students from one grade to the next. The approach using only students with scores in both years of the comparison is commonly referred to as a *longitudinal model*. It has the appeal that the school is only held accountable for students who were in the school for the period between the first and second test administrations. Although this feature of the longitudinal approach may seem an advantage to schools, it has the clear disadvantage of excluding mobile students who change schools from one year to the next and students who for some other reason are tested in only one of the years being compared. Therefore, the educational system is not held accountable for these students.

Table 2. Criteria for School Performance Grades

Minimum criteria for school performance Grades C, D, and F				Higher performing criteria for school performance Grades B and A			
	FCAT Reading	FCAT Math	Florida Writes!		FCAT Reading	FCAT Math	Florida Writes!
Elementary	60% score level 2 & above	60% score level 2 & above	50% score 3 & above	Elementary	50% score level 3 & above	50% score level 3 & above	67% score 3 & above
Middle	60% score level 2 & above	60% score level 2 & above	67% score 3 & above	Middle	50% score level 3 & above	50% score level 3 & above	75% score 3 & above
High	60% score level 2 & above	60% score level 2 & above	75% score 3 & above	High	50% score level 3 & above	50% score level 3 & above	80% score 3 & above

Note. Grade description criteria quoted from Florida Department of Education (1999, p. 1).

An alternative that avoids this disadvantage is to base the accountability on a comparison of the performance of all students in the school in, say fifth grade in 2000, with that of all students in that school who are tested in the sixth grade in 2001. This approach has been called a *quasi-longitudinal* approach. It has the advantage that all students in the school at the selected grades influence the results in a given year.

Both the longitudinal approach and the quasi-longitudinal approach require comparable tests across each grade

Table 3. Williams School
Colorado Student Assessment Program Score 3rd-Grade Reading Proficiency Levels

Year	Unsatisfactory	Partially proficient	Proficient	Advanced	Proficient or above	No scores	Total students
1998	13.13	25.25	58.59	3.03	**61.62**	0	99
1999	12.61	20.17	62.19	3.36	**65.55**	1.68	119

Note. Scores are percentages of students at that level.

level compared. Both require annual testing in every grade used in the accountability system and are generally associated with the use of off-the-shelf tests or measures with characteristics similar to such tests. A downside to off-the-shelf tests is that they are usually not matched to state content standards. Because schools usually align instruction to the content that they are tested on, the state standards often become a lower priority.

North Carolina is an example of a state that uses the quasi-longitudinal approach in its "ABC" school accountability system. At least a year's worth of growth for a year of schooling is expected.

North Carolina uses the average rate of growth observed across the state as a whole from one grade in the spring of 1993 to the next grade in the spring of 1994 as a benchmark against which improvement for students in a given grade in one year to the next grade the following year is judged. Comparisons to expected growth are used to classify schools into one of four categories: exemplary schools, schools meeting expected growth, schools having adequate performance, and low-performing schools (Table 4).

The Tennessee Value-Added Assessment System (TVAAS) is perhaps the best-known and most often cited state accountability system that relies on matched student-level longitudinal data for reporting of school, district, and teacher performance. Developed by William L. Sanders (see, for example, Sanders & Horn, 1994; Sanders, Saxton, & Horn, 1997), TVAAS uses sophisticated data-analysis methodology that allows the use of gains in student achievement from one year to another as the basis for holding teachers, schools and districts accountable. Student achievement data from several previous years are used as the basis for estimating gains in a particular year.

Adjusting School Rankings for SES

Research has well documented the overriding effect of socioeconomic status (SES) on student achievement. To account for this SES factor, some states such as California and Pennsylvania report "similar schools scores" to supplement their regular school rankings. The use of background measures is controversial, however, because they imply a lower set of expectations for less affluent students. They may also mislead educators and policymakers to presume that schools are doing better than they really are or not as well as they really are.

Table 5 shows recent rankings in a California district for five elementary schools to which we have given fictitious names. Of particular note is the ranking for Ash Elementary School. Although its overall 2000 rank of 6 is lower than the ranks of the other four schools, Ash has a similar schools rank of 9, higher than all other elementary schools in this district. Many other California district scores show similar disparities between the two different school ranking methods, creating a fair amount of confusion among educators and the public. Some educators have seized on the similar schools rankings, saying that their schools perform very well when compared to other similar schools, but failing to mention the schools' lower results on most performance measures.

Table 4. North Carolina ABCs Results for All Schools 1999 - 2000

Award or recognition category	K-12	Percent
Schools Making Exemplary Growth / Gain	959	45.3
Schools Making Expected [not exemplary] Growth / Gain	514	24.3
Schools Receiving No Recognition	597	28.2
Low-Performing Schools	45	2.1
Total Schools	2115	99.9[a]

[a] Percents do not total 100 due to rounding off.
Note. ABCs data quoted from North Carolina Department of Public Instruction (2000, August).

Table 5. School Academic Performance Rankings From a Sample California School District

School	2000 API	2000 Rank	Similar school	Target API
Ash Elementary	697	6	9	702
Birch Elementary	721	7	2	725
Carnation Elementary	742	7	2	745
Delta Elementary	731	7	3	734
Elm Elementary	776	8	3	777

Accepting or indeed promoting similar schools rankings may lead to lower expectations for students from different backgrounds. Because of the link between socioeconomic status and ethnicity, reducing the expectations for students from low SES backgrounds typically means lower standards for African American and Hispanic students.

Recommendations

No school reporting method is without some disadvantages; however, the recommendations below may enhance the likelihood that assessment systems will contribute to the overarching goal of improving student learning, while minimizing some of the potential negative effects. Although several suggestions repeat an earlier CRESST brief, they bear repetition in our high-stakes accountability environment.

1. **Place more emphasis on school improvement than on current performance.** This allows for differences in starting points while maintaining high standards and expectations of improvement for all.
2. **Report the margin of error for any school result.** All measurement systems, including polls, surveys, and even scientific tests, contain some degree of error. To avoid improper use of test scores, states should report the probability that a student or school has been misclassified.
3. **As required by the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), evaluate the validity of the uses and interpretations of assessment results.** Validate the *full* accountability system including standards, tests, alignment, professional development, rewards, sanctions, teaching quality, curriculum, and resources in addition to the positive and negative effects.
4. **Validate trends with results from other indicators** such as the National Assessment of Educational Progress, other state tests, and results from college admissions and placement tests such as the ACT, the SAT, and Advanced Placement tests.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bush, G. W. (2001). *No child left behind. A blueprint for education reform.* Washington, DC: U.S. Department of Education. Available March 21, 2001: http://www.ed.gov/inits/nclb/index.html

Florida Department of Education. (1999). *School accountability report guide: June 1999.* Available March 16, 2001: http://www.firn.edu/doe/bin00018/guide99.htm

Linn, R. L. (1998). *Standards-based accountability: Ten suggestions* (CRESST Policy Brief). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Linn, R. L. (2001). *The design and evaluation of educational assessment and accountability systems* (CSE Tech. Rep. No. 539). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Linn, R. L., & Baker, E. L. (1999, Spring). Standards-based accountability systems' adequate yearly progress. Absolutes, wishful thinking, and norms. *CRESST Line,* pp. 1-7.

North Carolina Department of Public Instruction. (2000, August). Number and percent of schools receiving awards and recognition 1997-2000. In *A report card for the ABCs of public education. Growth and performance of North Carolina schools, 1999-2000.* (Vol. I). Available 16 April 2001: www.ncpublicschools.org/abc_results/results_00/1997_2000.html

Orlofsky, G. F., & Olson, L. (2001, January 11). The state of the states. *Quality counts 2000* (Vol. 20, No. 17, pp. 86-88). Washington, DC: Education Week.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8,* 299-311.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee Value-Added Assessment System: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press, Inc.

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed in this publication do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement or the U.S. Department of Education. To order copies of this Policy Brief, contact Kim Hurst, 310-794-9140, email: kim@cse.ucla.edu, or write to Kim at CRESST/UCLA GSE&IS Building, Mailbox 951522, Los Angeles, CA 90095-1522.