

PREL BRIEFING PAPER

November 2000



Pacific Resources for Education and Learning
1099 Alakea Street ■ 25th Floor ■ Honolulu, Hawai'i 96813
Phone: (808) 441-1300 ■ Fax: (808) 441-1385
E-mail: askprel@prel.org ■ Website: www.prel.org

Assessment and Accountability

By Don Burger*

Accountability and assessment are terms that are frequently misused and misunderstood. This document is intended to stimulate conversations about assessment of and accountability for student learning, to raise assessment issues that may prevent improved student learning, and raise awareness about the relationship between assessment and accountability. Different assessment tools and strategies for establishing and maintaining accountability are outlined. While the main topic of this document is assessment and accountability, curriculum and instruction are integral parts of the education process and will be addressed as appropriate.

Definitions

The term “assessment” will be used to describe the process of determining what students know and how well they can apply it. Assessment involves making a judgment; to make assessments, educators rely on a variety of tools, including tests, homework, participation, projects, and other indicators. The term “test” will be used to describe an activity in which students demonstrate what they know and what they can do.

Test formats include paper and pencil exams, essays, performances, and quizzes. There are a variety of possible response formats, including multiple-choice, true/false, matching, or constructed response. Students may be asked to fill in a blank or to write a sentence, paragraph, or essay. For standardized testing, all students take the same exam under the same conditions. For non-standardized testing, students are invited to demonstrate their learning in a number of different ways.

Tests can be norm-referenced, in which case scores are compared to a national sample; or criterion-referenced, in which case student performance is measured against preset criteria. (Standards-based tests are criterion-referenced tests.) With all of these options at their disposal, policy makers must make informed choices.

* Dr. Donald Burger is the Director of the Pacific Assessment Systems and Services (PASS) Program at PREL.

Assessment and Accountability

Traditionally, classroom assessment and school/district/state accountability assessment have been separate and exclusive processes. The purpose for each type of assessment is different. The purpose of classroom assessment is to provide teachers with the necessary information to gauge student progress and assign report-card grades. Classroom tests are designed to measure how well students learned what was taught. Teachers record grades for a variety of student work, including tests, homework, and projects. The teachers assess all their information and calculate an overall grade for each student. By the end of the term or school year, the teacher should have a clear understanding of what each student knows and can do.

Accountability tests are designed to give school leadership and policy makers a means to evaluate the effectiveness of the system's curriculum and instruction. Accountability tests are typically given at a single point in time during the school year, usually in the spring. Standardized norm-referenced achievement tests (NRTs) are popular accountability tests because they compare local students against a national sample. Typically NRTs have been entirely "bubble tests," or tests with machine-scored multiple-choice response formats. Machine-scored tests are relatively inexpensive to use; however, they are only effective if the content tested corresponds to the content taught by the classroom teachers. The match between what is taught and what is tested is known as "alignment."

Schools, districts, or states purchase an NRT knowing that the alignment varies from grade level to grade level. For example, if fifth-grade reading scores are lower than fourth-grade reading scores, this may in part be due to different degrees of alignment. In other words, some students might have been tested on material they had not been taught. Because alignment affects accountability reporting, trend-line comparisons are made at specific grade levels rather than by following a cohort, or group of students, across several grade levels.

The separation of assessment and accountability has created distrust and animosity among teachers, principals, and central office administrators. Teachers frequently discount accountability tests like NRTs for a variety of reasons. In addition to problems with alignment, feedback from the test is delayed and it is difficult to translate the test data into changes in instructional strategies. Results for spring testing, administered in March or April, are often returned just prior to the end of school year. The last few weeks of school are packed with closing activities and teachers have little time to review test results.

On the other hand, administrators and policy makers discount teacher assessments because they are liable to subjectivity, grade inflation, and poor test design. The result has been assessment and accountability processes that are entirely independent of each other and that don't result in continuous improved student achievement. The assessment and accountability systems are dysfunctional.

Implications for Policy Makers:

- Establish clear purposes for classroom and accountability assessment.
- Create specific content standards for students, teachers, administrators, and schools.
- Conduct an alignment study of your accountability test at each grade level and in each content area tested.

Leveraging School Improvement

Unhappy with student-learning trend data over the past ten years, administrators, school boards, and legislators have been trying new approaches to improving student learning. Many states are constructing accountability assessments aligned with agreed-upon content standards. Kentucky, Maryland, Connecticut, Vermont, Colorado, and many other states are working on standards-based state accountability assessments. Several states, including Kentucky, Maryland, and Minnesota, are using tests to leverage the system to improve student learning.

One method of leveraging school improvement is to attach high stakes to state test results, an approach employed by Maryland, Minnesota, and others. In high-stakes assessment situations, incentives or consequences are tied to test scores. Consequences can affect students, teachers, administrators, or the system. For teachers and principals, “high stakes” may mean that they can be involuntarily moved, or in some cases, fired. For students, “high stakes” may mean that a single test score determines who receives a diploma, who graduates, or who moves to the next grade level. Every initiative, however, has both intended and unintended outcomes. One unintended outcome of high-stakes assessment is cheating. Teachers and administrators are cheating on high-stakes tests (Kantrowitz & McGinn, 2000; Maryland Department of Education, 1999; Thomas & Wingert, 2000).

Implications for Policy Makers:

- Using a single test to leverage improved student learning sounds attractive but can carry major negative consequences for students, teachers, administrators, and the educational system.
- High-stakes assessment should be approached only after careful consideration and planning. Most school systems are not ready to deliver on “high stakes” accountability. On the other hand, some schools will never deliver on the promise that “all students can learn” without deadlines and consequences.

Measuring Classroom and Accountability Systems Effectiveness

Trend-line charts have two valuable properties. First, trend-line charts are excellent tools that describe whether or not student learning is improving. This is an important indicator of system quality and equity. Second, trend-line charts are a good indicator of how well the curriculum, instruction, assessment, and accountability systems are working together. The trend line should reflect continuous improvement evidenced by increasing scores, mean-percentile-ranks scores on NRTs, and an increasing percentage of students at or above standards on standards-based tests. If not, something in the system, whether curriculum, instruction, assessment, or accountability, is broken and needs to be redesigned.

National Assessment of Educational Progress (NAEP) tests provide excellent trend-line charts. NAEP tests are criterion- rather than norm-referenced. They are designed to provide indicators of student learning relative to fixed performance standards. The NAEP scale scores are then plotted as trend lines (see Chart A below).

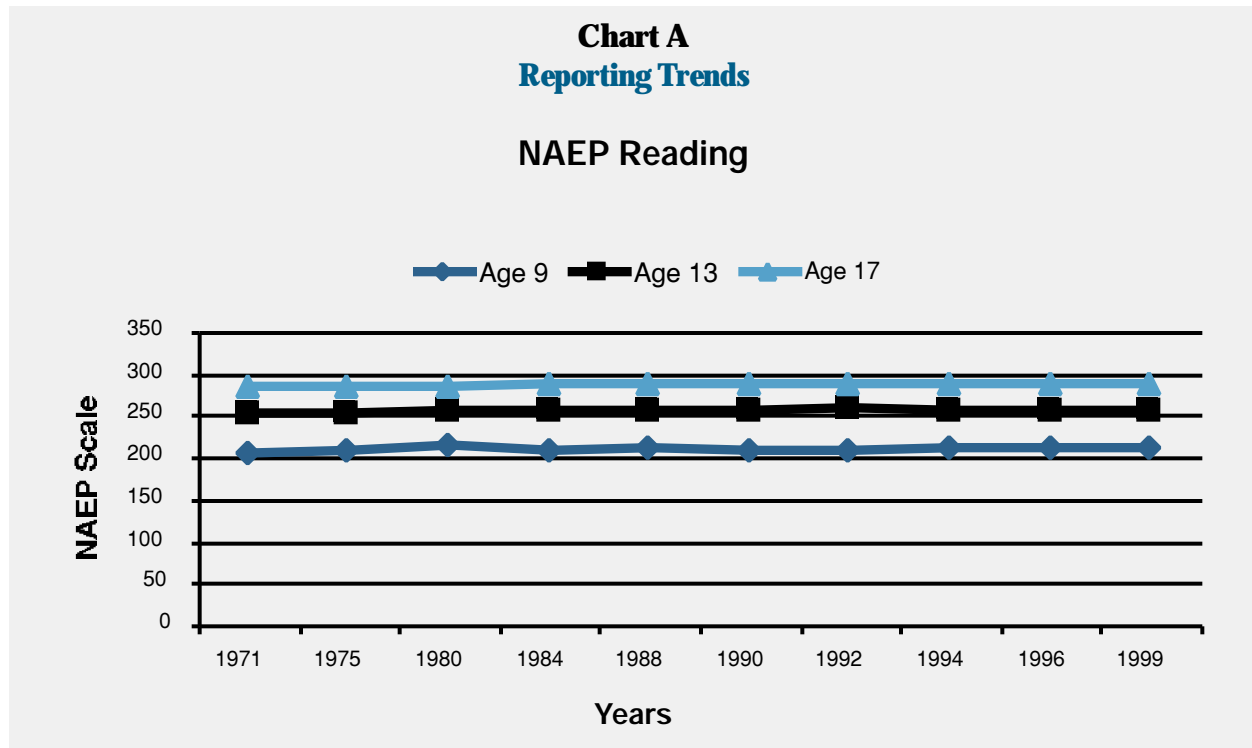


Chart A shows no change in student-reading performance in over 30 years; there is no evidence of improvement. Is the NAEP test part of an integrated curriculum, instruction, and assessment system? By design, it is not. This data suggests that improved student learning does not result simply from administering a national standards-based test.

Implications for Policy Makers:

- Use data to create conversations about school improvement and the quality and equity of local education.
- Trend-line charts are excellent tools to examine school improvement, quality, and equity.

Effective Reporting Tools

Chart A is easy to read and understand. However, the NAEP Scale, although statistically accurate, is difficult to explain. NAEP also reports data to the public using performance levels like “Advanced,” “Proficient,” “Basic,” and “Below Basic” (see Chart B). Charts using descriptor categories rather than numbers are easy to interpret and easy to understand. Each term describes a range of student skill and knowledge rather than a single score. For example, “Advanced” represents a superior performance for students in the grade level tested. “Proficient” represents solid academic performance. “Basic” denotes partial mastery of prerequisite knowledge and skills fundamental for the grade level. “Below Basic” is reserved for those students not yet able to demonstrate any of the prerequisite skills. The NAEP website includes examples of student work for each performance level (<http://nces.ed.gov/nationsreportcard/site/home.asp>).

The chart below (Chart B) answers the question, “What percentage of students score in each of the four performance bands?” Chart B shows that the percentage of students scoring in the “Proficient” or “Advanced” categories is increasing slightly. It also shows that nearly three-fourths of the students score in the “Basic” or “Below Basic” range.

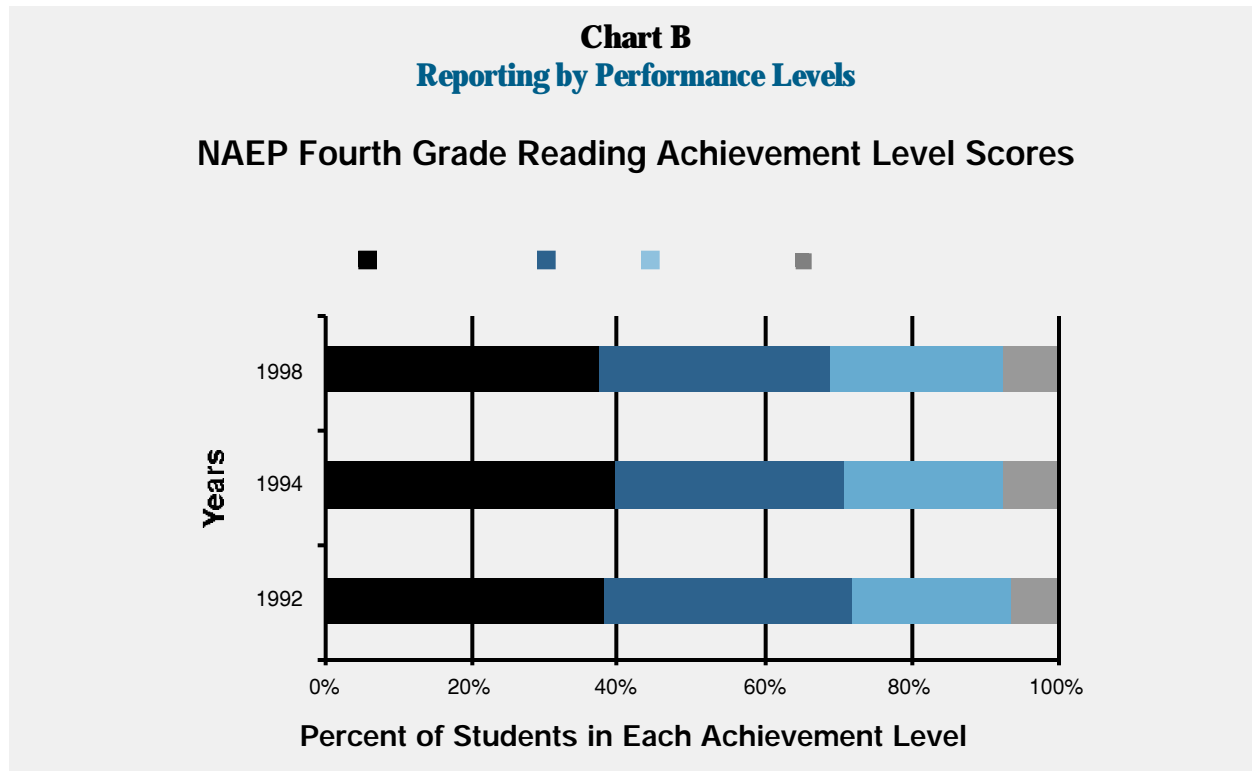
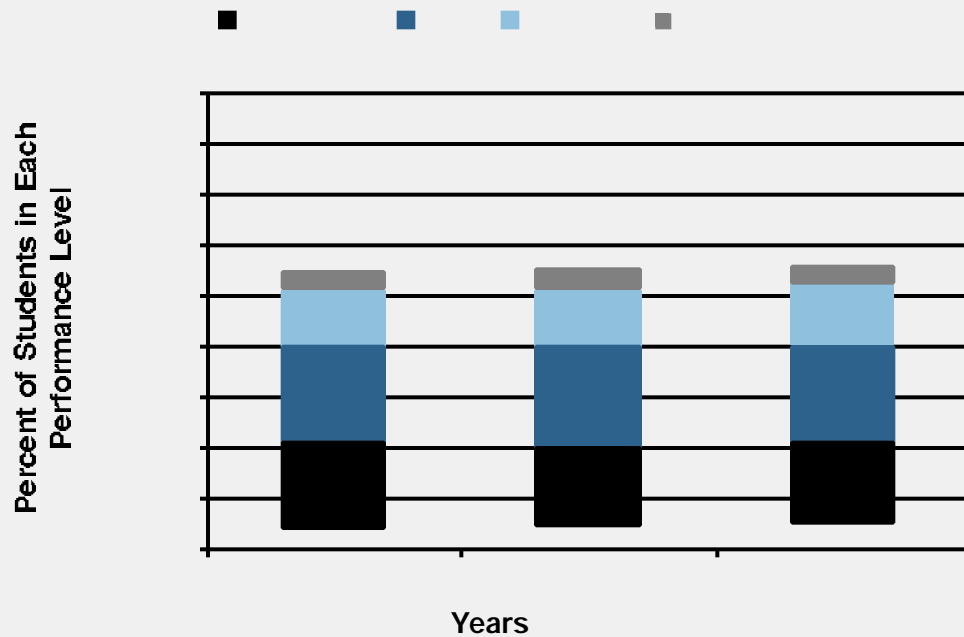


Chart C illustrates another reporting version of the NAEP data. Here the “0” line separates “Advanced” and “Proficient” scores from “Basic” and “Below Basic Scores.” Students scoring above the “0” line meet or exceed the standards, while students below the “0” line have not yet met the standards. The minus number below the “0” indicates the percentage of students scoring in the “Basic” and “Below Basic” categories. In Chart C, it is easier to determine that only 30% of the students have met the standard.

It’s a good idea to ask parents and community leaders to assist in determining which data presentation most clearly and correctly describes the data. Administrators may wish to create several data displays of the same information and ask school community members to interpret the results. Their interpretation will identify the format that works best for the school community.

Chart C Reporting by Performance Levels

NAEP Fourth Grade Reading Test



Implications for Policy Makers:

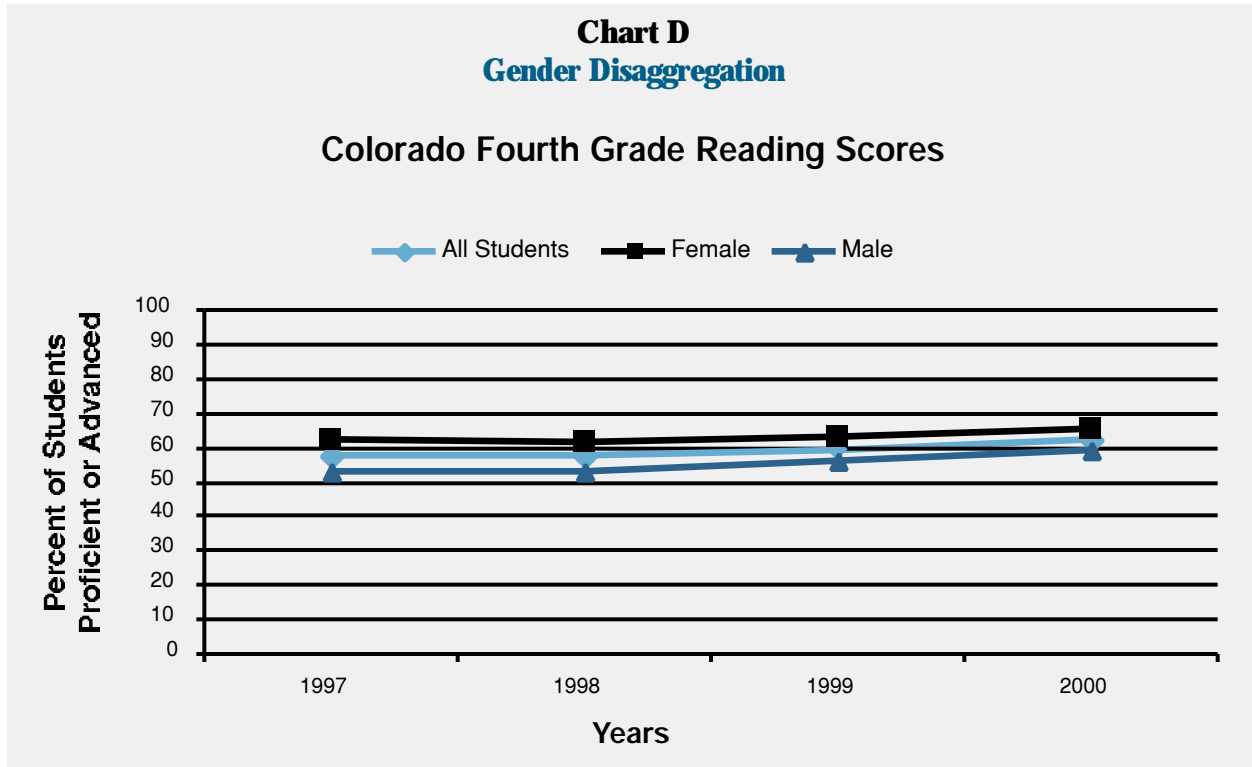
- Use data to create conversations about school improvement and the quality and equity of local education.
- Employ effective tools like trend-line charts to examine and report school improvement and educational quality and equity.
- Ask school community members to assist in selecting reporting formats.

Data Disaggregation

Many school districts and states categorize data to gain additional information about the equity of public education. This technique is known as “data disaggregation.” The state of Colorado does an excellent job of separating and reporting data by district size, gender, race/ethnicity, handicapping condition, accommodation, program, time in district, and time in school. The Colorado Department of Education website (http://www.cde.state.co.us/index_assess.htm) is a wonderful assessment resource for a variety of Colorado achievement data.

Chart D uses trend lines to examine educational equity by separating fourth-grade reading data into two groups: female and male. It attempts to answer the question, “Does the educational system produce equitable results?” by comparing the percentages of boys and girls who are reading at “Proficient” or “Advanced” levels. Trend lines in Chart D show that the percentage of girls scoring at “Proficient” or “Advanced” levels is consistently higher than the percentage of boys scoring at

those levels, and that the scores of both girls and boys are improving slightly. Disaggregated trend-line charts can help educators determine where attention should be focused.

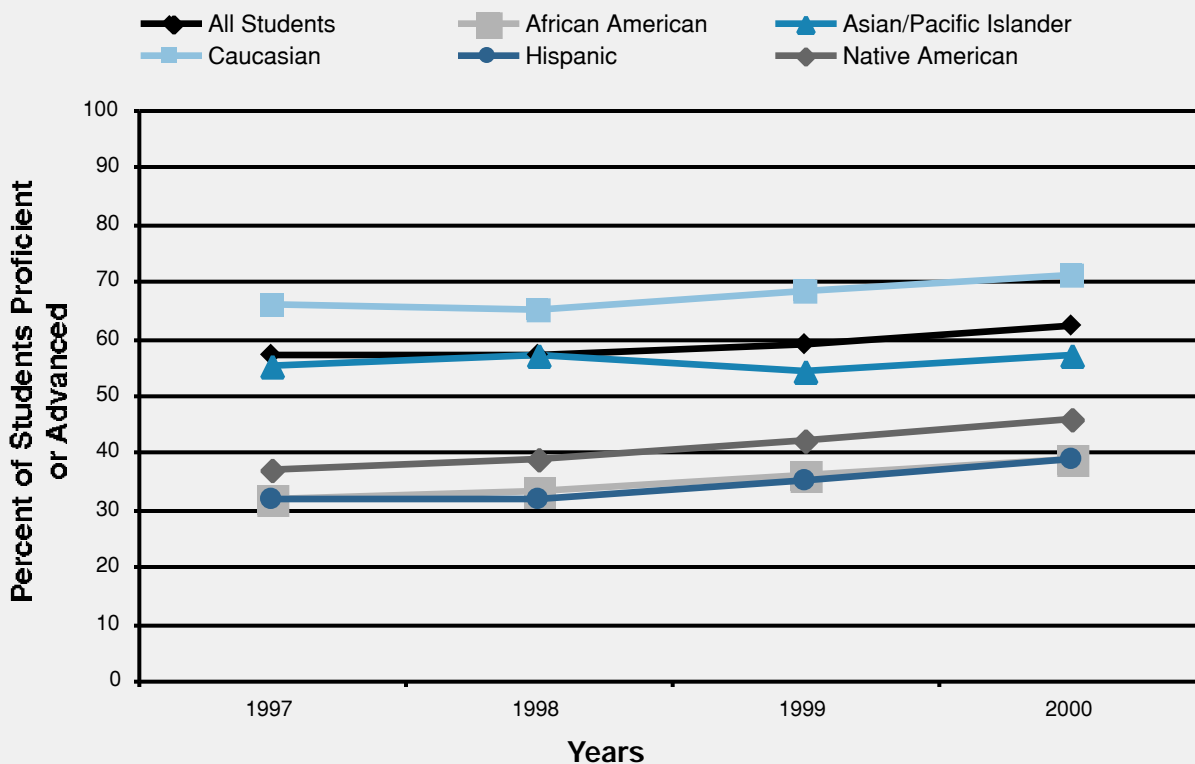


The data show that the percentage of girls and boys reading at or above the standard is nearly equal: Overall performance of boys and girls is approximately 60% at or above the standard. While these performance ratings in and of themselves may not be acceptable, the data suggest that the system yields equitable results by gender.

Disaggregation by ethnicity is used as an indicator of system equity, and Chart E uses trend lines to compare equity among ethnic groups on the Colorado Fourth Grade Reading Test. Chart E indicates that White and Asian students fare best in the Colorado Educational System, followed by African American and Hispanic students. Native American students fare least well in the system. The data point to areas where educators in Colorado should concentrate resources to improve reading. Trend-line charts are an excellent way to demonstrate equity, or lack of equity, in the system.

Chart E Race/Ethnicity Disaggregation

Colorado Fourth Grade Reading



Implications for Policy Makers:

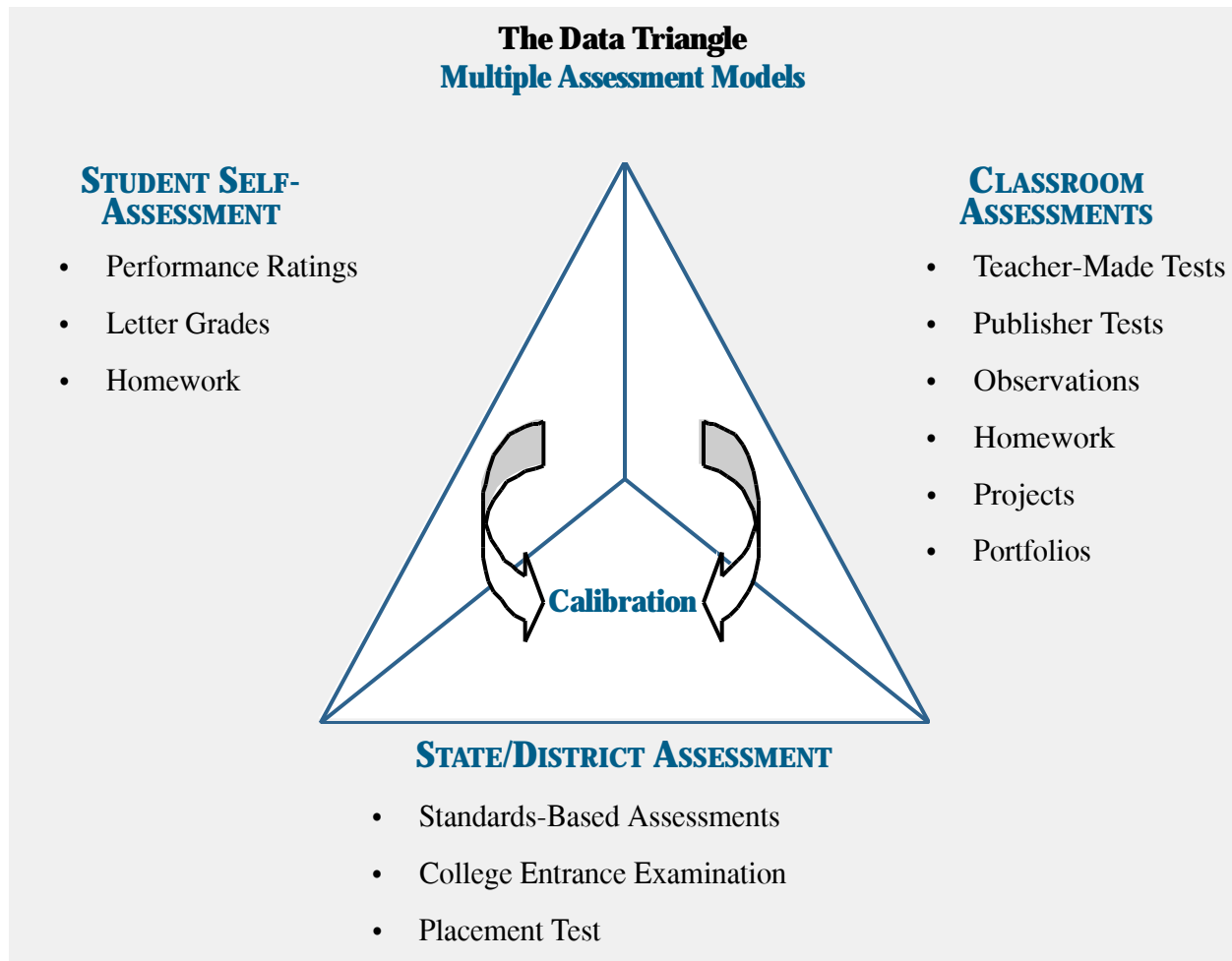
- Use data to create conversations about school improvement and the quality and equity of local education.
- Trend-line charts are excellent tools to examine school improvement, quality, and equity.
- Separating (disaggregating) and reporting data by demographic variables such as gender and ethnicity sharpens the focus of the assessment and accountability systems. Disaggregation can demonstrate the effectiveness of efforts to improve equity.
- Consistently measure and publicly report those factors the system values.
- Report to the public using tools that are easy to understand and interpret correctly.

Integrated Assessment Systems

A meta-analysis by Black & Wiliam (1998) reports that aligning classroom-assessment content standards has the greatest effect of all the approaches currently in use in the schools. They report “effect sizes” in the range of 0.4 to 0.7. “Effect sizes” are methods researchers use to evaluate the impact of an innovation. An effect size of 0.4 would increase the scores of students by 15 percentile points on a norm-referenced test or precipitate grade equivalent gains of 1-3 years. An effect size of 0.7 translates into score improvements of approximately 45%.

The standards movement offers an opportunity for schools/districts/states to design complementary, congruent, and integrated assessment and accountability systems. Standards provide the opportunity for school communities to reach consensus on what students should know and can do. Curriculum, instruction, classroom assessment, and accountability assessment need to be aligned with the content standards. Alignment presents an opportunity to link classroom and accountability assessment.

Integrated assessment systems combine academic information from three sources: students, teachers, and district/state assessments (see Data Triangle). Creating excellent student self-assessors is just as important as creating student learning. Students should know and be able to perform those tasks agreed upon by the community and know how well they perform them. Teachers must be able to create classroom assessments that are of sufficiently high quality that they can be reported along side the district/state accountability test.



Integrated assessment systems provide multiple sources of information that have been aligned with content standards. Each of the assessment components is a valid measure of the standard and each has performance levels that are similar in rigor. Once calibrated, student and teacher data can be reported alongside state standards-based data.

Implications for Policy Makers:

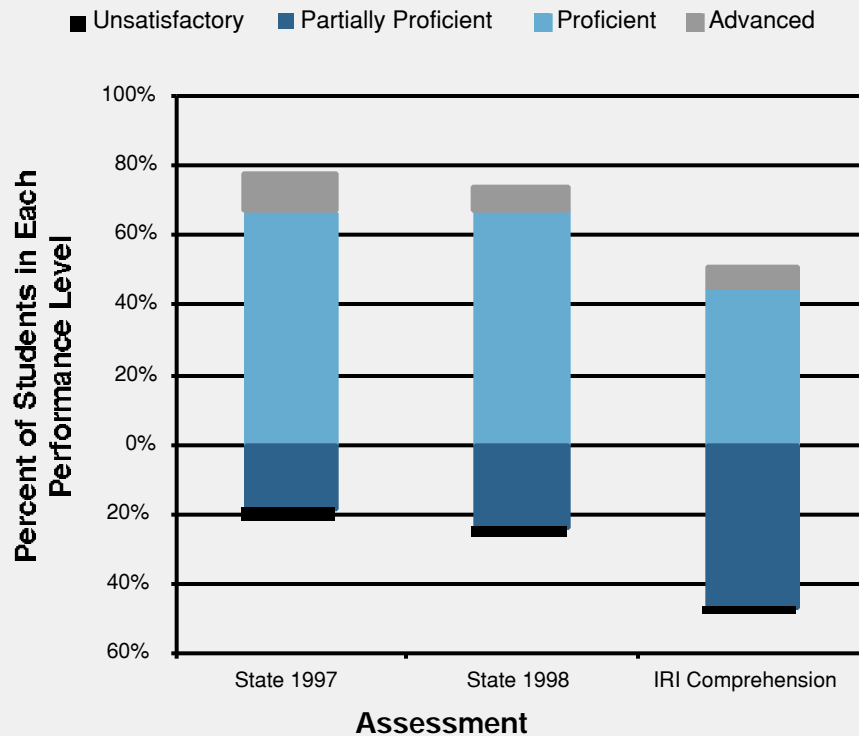
- Continuous professional staff development in assessment is necessary to improve the quality of classroom assessment. The research shows that when performed properly, excellent classroom assessment practices will have the largest impact on student learning. This effectiveness is an outcome not of assessment by itself, but of a system that incorporates integrated assessment, curriculum, and instruction.

Calibrating Classroom Assessment

Classroom assessment must also be calibrated to the state’s test so that both scoring systems are equal in difficulty. Calibrating tests is one way to make sure that teachers and the state agree on what makes a “Proficient” or “Advanced” performance. Teachers can then predict accurately on any day how their students would fare on the state test. Chart F shows the calibration of a classroom individual reading inventory (IRI) and a state reading test. The IRI scores are reported as instructional levels ranging from 1.0 to 11.0 including half grades. A score of 2.5 places a student at instructional level two and a half.

Chart F
Fourth Grade Reading Calibration

Calibration of Individual Reading Inventory and State Fourth Grade Reading Scores
(Conversion Table: 1=1.0-2.5; 2=3.0-4.5; 3=5.0-8.5; 4=9.0-11.0)



The IRI calibration that is the closest fit with the state reading test is the State Reading Test and Individual Reading Inventory Calibration (see below).

State Fourth Grade Reading Test & Individual Reading Inventory Calibration Table	
State Performance Level	Individual Reading Inventory Instructional Levels
Advanced	9.0 – 11.0
Proficient	5.0 – 8.5
Partially Proficient	3.0 – 4.5
Unsatisfactory	1.0 – 2.5

Another use for data is to challenge current beliefs and assumptions. The tables show that the instructional levels are less demanding than the expectations the state sets for fourth-grade students in reading. To meet the reading standard, students must perform at the 5.0 instructional level. Teachers would then use the 5.0 instructional level as the goal for fourth-grade students. Anytime a teacher gives the IRI, she/he will learn how his/her student would have done on the state test if given that day.

Implications for Policy Makers:

- Use data to challenge assumptions and beliefs about students, programs, and learning.
- Create the opportunities teachers, administrators, and board members need to learn about assessment and accountability.
- Provide continuous professional staff development on assessment design and interpretation for teachers and administrators.
- Provide state-standards-based assessments against which teachers and students can calibrate classroom assessments.

References

- Black, P. & Wiliam, D. (1998, October). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 139-148.
- Kantrowitz, B., & McGinn, D. (2000, June 19). When teachers are cheaters. *Newsweek*, 48-49.
- Maryland Department of Education. (1999, June). *Summary findings of allegations of cheating on the 1998 Maryland State Assessment*. Presentation at the CCSSO Annual Assessment Conference, Snowbird, UT.
- Thomas, E., & Wingert, P. (2000, June 19). Bitter lessons. *Newsweek*. 50-51.