measured progress

Proficient, Eligible to Graduate, College-Ready?

The mystery of achievement-level assessment results

Stuart Kahl | October 2015

Years ago in large-scale testing programs, students received test scores and corresponding percentile ranks. A test score was a direct reflection of how many points a student earned, and a percentile rank represented the percentage of test-takers the student tied or outscored. Neither statistic communicated how well a student knew or was able to do what was expected at that point in school. Nor did average scores and median percentile ranks communicate the same for groups of students. The solution: achievement-level reporting.

The Double Meaning of "Standards"

Notice I said "achievement-level reporting" instead of "standards-based assessment." The latter is accurate, but its double meaning is the cause of a great deal of confusion, which has led people who are somewhat lacking in assessment literacy to mislead policy makers and the general public. The term "standards" in the world of educational assessment refers to two distinct things: 1) the content domain the test items sample and 2) how well the test-takers must perform to be designated as achieving different levels of proficiency. The failure to distinguish between these two meanings has prevented many from understanding why two tests measuring similar, or even identical, content standards can produce very different results in terms of "percentages of proficient students."

This problem existed over two decades ago when the National Assessment of Educational Progress (NAEP) introduced achievement-level reporting for state results, it has continued through the No Child Left Behind (NCLB) era, and it continues today as states ponder the future of their assessment programs. This paper explains the problem and tracks its history.

> "The term 'standards' in the world of educational assessment refers to two distinct things."

Standard Setting

State accountability assessment programs use a process called "standard setting" to establish "cut scores." A cut score is the minimum score a student must earn on a test to be categorized in a particular achievement level. The process involves panels of educators and non-educators making judgments about test items or samples of student work, which when aggregated and analyzed lead to the desired cut score(s). Often, three cut scores are created that define four performance levels, and often the second highest level is called "Proficient."[1] The cut scores that emerge from the work of the standard-setting panels are ultimately presented to a policy-making body that accepts, rejects, or sometimes adjusts the cut scores.

[1] Of course, NCLB initially required/expected that after 14 years, almost all students in every school would score at the proficient level or higher on the state assessments, and "percentage proficient or above" became the primary focus of assessment program reporting.

State NAEP and Individual States' Proficiency Results

NAEP used achievement-level reporting in the early 1990s when it conducted its "trial state assessments." States in the Southeast Regional Education Board (SREB) participated in this trial program, which ultimately became the vehicle to make legitimate comparisons of states in terms of their students' academic achievement. At that time, however, it was quite puzzling to many why the percentages proficient from NAEP differed significantly from percentages proficient reported by the individual SREB states' own assessment programs. How could testing programs supposedly measuring the same or similar knowledge and skills come up with such different results? (Hint: See Standard Setting, above.)

NAEP set some pretty high cut scores for proficiency in reading and math, typically identifying only a quarter to a third of students as Proficient or higher. Many states have reported much higher percentages based on their own assessments. One common but mistaken conclusion was that the states' own tests were too easy or not as rigorous as NAEP tests. In fact, which tests were easier or harder was irrelevant—**the primary factor leading to the reported achievement-level results was where the cut scores were set**. If one computes the correlation coefficient representing the relationship between the test scores (not achievement levels) from two general achievement measures in a subject, a fairly high correlation is obtained.

> "What matters is **not** whether different tests yield similar proficiency results. What matters is that a state's assessments are equated from year to year and that the cut scores are held constant."

To assessment-savvy folks, differing percentages proficient on NAEP and state tests were not a problem. Why should the results be the same? After all, the states' programs might well have had different primary purposes: school program improvement, identification of students needing remediation, advancement to the next grade, graduation, and even perhaps college readiness. And they most likely used different standard-setting procedures, which have been shown to produce different results. What matters is **not** whether different tests yield similar proficiency results. What matters is that a state's assessments are equated from year to year and that the cut scores are held constant, so that decisions based on test results are fair and so that change (hopefully, improvement) can be monitored. That results across individual states' tests could not be compared was not a problem, since NAEP was ultimately designed to allow such comparisons. "State NAEP" became a measuring stick for all states by 1996.

State Assessments and Commercial Tests

A few years later, two New England states wrestled with this same issue of test comparability. In New Hampshire, two university reading educators wrote a scathing report claiming the state assessment results were not trustworthy because they were not correlated with those of a commercially available test, the Iowa Test of Basic Skills (ITBS). Had these "researchers" correctly computed a correlation of scores on the two measures, they would have found the usual high correlation. Their false conclusions were based on their misunderstanding of achievement-level reporting and of where the two different programs established their cut scores. The ITBS identified almost three-quarters of students as Proficient or above, while the New Hampshire program used cut scores that put approximately a third of its students in the upper two categories.

Massachusetts fared better with a similar situation. In the early years of the Massachusetts Comprehensive Assessment System (MCAS), third graders took the ITBS reading test, while fourth graders took MCAS. There was serious concern about the fact that

parents could be told their children were proficient in third grade, then be told that was not the case in fourth grade, even if the students experienced typical "growth" during the interval between the two tests. Possibly because MCAS communication was particularly effective on this matter, a significant public relations issue was avoided.

These anecdotes show how mixed messages can be sent from different testing programs and why it is important that achievement-level reporting be better understood by consumers of test results. However, even if exact comparability across state programs cannot be accomplished, the revelation that some states set cut scores substantially lower than others might suggest the need for "raising the bar" in some states. After all, as the research indicates, expectations for students do influence their motivation to learn and, ultimately, their achievement.

The "Common Core" Era and Comparability

While this discussion so far might seem to be about ancient history, the problem has not gone away, and probably will persist as long as achievement-level reporting is used. **However, this is not an argument against achievement-level reporting**. I believe the better information it provides on how well students are performing relative to content standards is especially important. Descriptions of what students at different levels tend and tend not to be able to do are useful. Nevertheless, the problem persists. At a launch event in D.C. for the Common Core State Standards (CCSS), one audience member stated that with states' adoption of the CCSS (aka Common Core), the academic performance of students in those states could now be appropriately compared. That person was surprised when I pointed out that common content standards were not enough, and that common or equated tests and the same cut points would also be needed.

It seems that a desire for comparability of test results across state programs was one of the driving forces behind the creation of the Common Core and the state assessment consortia, PARCC and Smarter Balanced. Yet with states dropping out of the consortia, and other barriers to state assessment comparability, it may be that policy makers will have to accept the lack of comparability across state tests and rely on State NAEP, which has provided that comparability for nearly a quarter-century. Besides, no matter where states set their cut scores, they can still identify high- and low-performing schools.

Assessment Program Choices for States

Currently, many states are deciding on the future directions of their accountability assessment programs. Failure to understand cut scores and performance-level reporting continues to be a problem. Officials in several states, in evaluating different program options, want to compare their reported percentages of students at or above certain levels. These statistics are not at all comparable—differences in results between programs are merely a matter of where the different programs happen to have set their threshold or cut scores for different achievement levels. Furthermore, cut scores for legacy programs, while they should be held constant for several years for monitoring changes in student performance, should be revisited periodically and can be re-set for a number of reasons, including new program purposes, provided the tests are measuring the "right stuff."

> "Assessment-literate educators and policy makers understand the difference between content standards and cut scores. . ."

The focus of the decision makers should be on what **content standards** the assessments address and how well they meet appropriate specifications for

content coverage and measurement quality. And if more than one potential program meets these content and quality requirements, then the next priorities should be other factors such as cost, local educator involvement/ownership, control over program changes, and the like.

Achievement Levels vs. Scaled Scores

Clearly achievement-level reporting has its advantages. Cut scores (and thus achievement levels) are useful for many decisions about students that tests are designed to inform. Nevertheless, this approach to the reporting of test results has limitations, too. In addition to understanding the lack of comparability of results across programs, it is important for consumers of test data to understand that information is lost when the full range of test scores is reduced to a few achievement levels. For example, two students, one scoring just at or above a cut score and the other just below it, would be indistinguishable in terms of their capabilities as reflected by their test scores, yet one could be labeled “Proficient” and the other not. In the same vein, two students scoring at the opposite extremes of the same achievement level could both be considered “Proficient” despite distinctly different capabilities. Assessment-literate educators and policy makers understand the difference between content standards and cut scores, the subtleties of achievement-level reporting, and whether scaled scores or performance levels are more appropriate for a particular use.