

# The Condition of Assessment of Student Learning in Arizona: 2005

Darrell Sabers  
*University of Arizona*

Sonya Powers  
*University of Arizona*

Reviewer: Thomas M. Haladyna  
*Arizona State University West Campus*

---

## ***Background***

There are at least five purposes for assessment of educational achievement. Four of these purposes are used as a framework for discussing the measures relevant to the assessment of educational achievement in Arizona. A fifth purpose—accountability—is addressed in *The Condition of School Accountability in Arizona: 2005*. This report will review the following assessment purposes: instructional guidance for teachers providing instruction in a particular subject area, comparison of student performance across subjects, indication of a school’s status within the state, and indication of a state’s standing in the nation. A discussion of teaching to the test and a consideration of how instruction affects test performance are addressed in the Findings section.

### **Purpose I: Instructional Guidance for Teachers Providing Instruction in a Particular Subject Area**

A teacher might use a student’s test score to help determine the level of instruction necessary for a student to progress in a subject. It is essential that the score yield information regarding how well a student performs in the subject area of interest.

## Purpose II: Comparison of Student Performance Across Subjects (Achievement Profiles)

This comparison may be made by the teacher, counselor, or parents who are interested in determining the strengths of the learner. The test scores in two or more areas of interest must be interpretable on the same scale, that is, the scores must be able to indicate the same level of performance if the student is equally proficient in the areas tested. A similar comparison may be desired to determine whether a class or larger group (school or state) has students who are more proficient in one subject than in another. For this comparison to be valid, it is necessary that equal scores reflect equal proficiency.

## Purpose III: Indication of a School's Status Within the State

This indication often involves a comparison of average scores on a test with a distribution of averages for other schools. Unfortunately, in the absence of a distribution of averages for other schools, a less accurate measure is often used where a school's performance is compared to the distribution of individual student scores. Another indicator of school status involves comparing the percentage of students achieving a certain level of proficiency (e.g., "meets or exceeds standards") with the distribution of percentages for other schools. These status indicators may also be used for measuring growth when available at different times for the same groups; however, that use of achievement tests is covered in *The Condition of School Accountability in Arizona: 2005*.

## Purpose IV: Indication of a State's Standing in the Nation.

This comparison is similar to Purpose III for schools except that the distribution of scores (averages or percentages) is for all the other states.

## ***Recent Policy Developments***

This section describes the types of scores that are used to report performance of students and schools and the development of standardized tests.

National standardized achievement tests were first developed based on the belief that states and school districts did not differ substantially in the general content of their curriculum guides, and the test blueprints were focused on the areas of agreement across these guides. It was expected that teachers would use state or local curriculum guides to determine what to include in classroom instruction and what to expect to be included in test blueprints.

## Test Score Types

Test results are reported using scores that can be interpreted without knowledge of the particular items that are included in the tests, because the test items are secure. The number of items correct on a test, sometimes referred to as a concept score, provides no basis for comparing performances of students taking different forms of a test or taking tests in different subject areas.

There are two types of scores commonly used: developmental scores and within-group scores.<sup>1</sup> The developmental scores are in the form of scaled (or scale) scores that can be used to show performance across different grade levels—a common example is the grade equivalent score (e.g., performing at eighth-grade level). Grade equivalents have fallen from favor because they are misleading due to the different meanings attributed to units of growth, for example, one month of growth represents very different amounts of learning at different grade levels. A more common developmental measure is a scaled score that has meaning only for a given test, but can be used to measure growth when accompanied by other information provided by the publisher.

Scaled scores can also be used for determining cut scores (cut-off points) to represent desired levels of performance such as “meets standards” or “exceeds standards” or for deriving other within-grade scores such as percentile ranks. A percentile rank describes the percentage of scores in a distribution that fall below a given score. The percentile rank is similar to the rank of the score in a distribution, except that typically the best score is ranked ‘one’ whereas the best percentile rank is 99 (100 and zero are not used for percentile ranks). Differences in percentile ranks do not represent equal

measures of a difference within a group, but other within-group scores are available when equal-interval scores are needed. Those equal-interval scores are not discussed in this report.

## **Norms, Norming, and Norm-Referenced Tests**

Providing meaning to a test performance requires referring a score to a distribution of scores from a group having taken the same test under the same conditions. These distributions of scores are called ‘norms’, and the process of obtaining the scores is called ‘norming’ a test. A sample of students in the nation is used to obtain a meaningful group for comparison for the national standardized tests, and these tests have become so closely identified with their scores referenced to norms that they are referred to as norm-referenced tests.<sup>2</sup> The percentile ranks reported on tests are considered objective measures of the relative achievement of students.

## **Standardized and Standards-Based Tests**

Levels of performance are a more subjective representation of achievement based on a state or national committee’s determination of how well a student should perform (e.g., to be considering mastering the subject matter of interest). These levels of performance are not usually emphasized (or even reported) with the national standardized tests, and the tests that do report performance levels are called standards-based tests to indicate that they are developed to report on mastery of content standards.

Some states purchased national standardized tests for statewide assessment, requiring all schools to give the same test. To better fit their state curriculum, states have more recently developed their own state assessments (often in collaboration with a test publisher). The No Child Left Behind legislation requires that states conduct annual testing to report on the progress of all students and schools, and to cooperate in the National Assessment of Educational Progress to monitor the learning of students in all the states. The results from any of these tests are reported using some type of comparison among states or schools—the common scores reported are within-groups measures such

as levels of performance (e.g., percentage of students who “meet or exceed standards”) or the percentile rank of the average score.

Arizona has used national standardized tests for decades as part of statewide testing programs. Now Arizona administers a standards-based test statewide. The tests currently used in Arizona are described below.

- Arizona's Instrument to Measure Standards (AIMS) is Arizona's annual standards-based assessment. Grades 3, 5, 8, and 10 have been tested in mathematics, reading, and writing; grades 4, 6, and 7 will be included starting in 2005. AIMS scores are reported as within-grade scaled scores, concept scores (i.e., number of items correct), and as levels of performance based on certain cut scores. The four levels of performance are: exceeds standards, meets standards, approaches standards, and falls far below standards.
- TerraNova (TN) will be used as a national standardized test for grades 2 and 9 starting in 2005. TN scores will be reported as National Percentile Ranks. Previously, the Stanford 9 was used in Arizona to provide National Percentile Ranks for students in grades 2 through 9.
- The above two tests are being incorporated into a Dual-Purpose Assessment (DPA) for grades 3 through 8. The DPA is intended to decrease testing time by including some items from AIMS and some items from TN in a single test for each grade level. Some items will function as both TN and AIMS items, and contribute to scores for both tests.
- The National Assessment of Educational Progress (NAEP) is a national standards-based test that has been used in Arizona and is now required by No Child Left Behind (NCLB) for grades 4 and 8 in reading and mathematics. States have an option to include testing of science and writing. NAEP performance levels are: advanced, proficient, basic, and below basic.

## *Findings*

Arizona's standing in the nation was a topic of much interest in the 2004 edition of this report.<sup>3</sup> A controversy arose because Stanford 9 and National Assessment of Educational Progress (NAEP) presented a different picture of the status of Arizona's students. An obvious difference was that the Stanford 9 results indicated that Arizona's students were making achievement gains whereas the NAEP indicated little or no gains. On the Stanford 9, Arizona students compared favorably with students in the nation, but on the NAEP they were below average when compared with students in other states. This year the Terra Nova (TN) will replace the Stanford 9 as the standardized achievement test for reporting the status of Arizona's students, but that does not ensure that there will be no discrepancy when NAEP and TN trends are compared.

Haladyna<sup>4</sup> suggested that the increase in Stanford 9 scores is reminiscent of Cannell's<sup>5</sup> "Lake Wobegon Effect" associated with states reporting pervasively above average test scores. The similarity of national standardized tests, both in content and methods of obtaining norms, results in similar long-term trends in student performance regardless of testing company. The trend indicates continuous improvement over consecutive years of administration of the same national standardized test while performance on NAEP appears to remain more consistent.<sup>6</sup>

The most publicized reason for the increase in test scores within a state is the possibility that teachers tailor their curriculum and focus to reflect the test's weighting of objectives. Along with teaching of testing skills, the conformity of curriculum to testing is often referred to as teaching to the test.<sup>7</sup> It is assumed that the curriculum becomes less comprehensive when students are "taught to the test," but a more important concern is that not all teachers place equal emphasis on teaching to the test, putting some students at a disadvantage. Also, because the national norms are not obtained from students who have been uniformly taught to the test, the differential focus on test preparation will distort the reported performance of students. At the extreme end, teaching *to* the test can become teaching *the* test, where teachers learn the specific test items that assess various

objectives, and teach those test items. Far fewer teachers use previous test scores to identify and remediate areas of weakness in their students, a practice that could be potentially more effective.<sup>8</sup>

Because of the increasing pressure on schools and states to perform, it is likely that teachers will feel pressure to use whatever resources they have available to produce student gains. Mehrens and Kaminski<sup>9</sup> describe seven points on a continuum ranging from teaching the curriculum without looking at what the test measures to having students practice on the same test they are to take in the testing program. If teachers chose the same point on this continuum for their instruction of all students, each student would have a fair chance to be measured validly by the test. What point should be chosen is a matter of debate. The reviewer of this report stated, “teaching to the test blueprint is teaching to the test. As the test is a sample of a larger domain, teaching to the test blueprint is an educationally unsound practice that leads to narrowing of the curriculum. Teachers should teach the Arizona content standards.<sup>10</sup>” That position on teaching to the test is quite conservative when compared to the following examples of guidelines available to educators outlining acceptable test preparation. The Standards for Educational and Psychological Testing state:

(Standard 3.20) The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. When appropriate, sample material, practice or sample questions, criteria for scoring, and a representative item identified with each major area in the test’s classification or domain should be provided to the test takers prior to the administration of the test or included in the testing material as part of the standard administration procedures.<sup>11</sup>

Likewise, the Code of Fair Testing Practices in Education (Revised) includes statement D1: “Inform test takers in advance of the test administration about the coverage of the test, the types of question formats, the directions, and appropriate test-taking strategies. Make such information available to all test takers.”<sup>12</sup>

Frameworks (or blueprints) and released/sample test items for Arizona's Instrument to Measure Standards (AIMS), TN, and the Dual-Purpose Assessment (DPA) are publicly available on the Arizona Department of Education (ADE) website.<sup>13</sup> CTB-McGraw Hill's website<sup>14</sup> also has information about the TN blueprints as well as a plethora of "teaching tools" marketed to help teachers teach to the test. Thus, it appears that ADE has taken a stance on teaching to the test by providing this information. The ultimate proof of the ADE position is that students have repeated opportunities to satisfy the graduation requirement of "meets or exceeds standards" on AIMS. Taking an alternate form of AIMS is the most extreme position on the continuum of test preparation except for the opportunity to practice on the actual test form. A high school senior taking the AIMS exam may have already practiced on several forms of the test, and the score might reflect practice effects in addition to proficiency in the area tested.

NAEP has been considered a more valid index of state achievement because of its low-stakes nature. Teachers and schools may have felt less pressure to improve performance on NAEP as compared to high-stakes tests such as AIMS. Although states may feel pressure to show relative improvement on NAEP, documentation of states encouraging districts or schools to improve does not exist.

It is not expected that teachers teach to a test that is not administered every year or in the same school every testing cycle. NAEP frameworks<sup>15</sup> are much more general and descriptive, and less instructional, compared to the other websites and supplemental materials. With the mandated state NAEP scores every two years for math and reading, it will be interesting to see if the increase in testing frequency and consistent participation of all states will lead to improvement in test scores in those subjects and not in optional subjects. It may be found that influences such as practice effects and teaching to the test that have affected national standardized tests will begin to influence NAEP as well, especially if NAEP becomes a criterion for state accountability.

It has become impossible to determine what is truly improvement in school, state, and national education. Burstein has suggested that more information on the background of test takers, including descriptions of the groups sampled for obtaining norms and the

test preparation allowed, be available with each test report. In addition, he suggested that “annual user norms” be used to report performance with respect to ‘new norm’ data. He warned “if we become too obsessed with measuring accurately the average performance of the students nationally, regionally, and locally, we may do a disservice to the educational improvement effort” because we corrupt the meaning of the measures.<sup>16</sup>

Reporting has been done poorly for school growth, especially because the Stanford 9 did not report “school norms.” Instead, the schools’ performances were reported using student norms (that is, comparing school averages to the percentile ranks for individual student scores in the nation). Because the TN does not have specified achievement levels, AIMS, using annual Arizona norms, would be the best source of information for comparing percentages of students at various performance levels across schools. Burstein’s recommendation that information in addition to the average score be included should be heeded.<sup>17</sup>

AIMS cannot be used for national status reports because these comparisons are not possible for tests developed and given in individual states. Even if ADE provides validity data for AIMS, a state test can never satisfy the need for a standard measure for national comparisons. Because states choose different national standardized tests and different items from test publishers that align with different state standards, NAEP is essential for describing state progress in the scope of the nation.<sup>18</sup>

NCLB mandates state assessments as the index of state achievement, but it also requires state participation in NAEP. This requirement may be a way to keep states honest about achievement gains. NAEP trends are more believable and more nationally accepted than trends found from state tests. The consistent participation of all states in NAEP will also yield informative long-term trends in state achievement. Although schools and districts may not be as concerned with NAEP progress as they are with their state assessments, national writers for *Education Week*,<sup>19</sup> *Education Trust*,<sup>20</sup> and the Manhattan Institute<sup>21</sup> rely on NAEP to rank and evaluate states. Other tests, like advanced placement tests, SAT, and ACT, that are accepted by national writers are taken

by students with special reasons for being tested rather than by comparable groups, and thus are not valid for comparing states.

Tests in Arizona have yielded very different pictures of performance of students within the state. Stanford 9 showed higher performance in mathematics than in reading, with an inflated measure of overall performance. Because more students met the Arizona standards in reading than in math, many readers of reports on AIMS data believed that Arizona students are less proficient in math than in reading despite the contradictory Stanford 9 results. Although Arizona students are consistently below the national average on NAEP in both reading and mathematics, it is interesting to note that Arizona has shown improvement in the mathematics portion of NAEP—a gain that may truly matter.

### *Policy Implications*

An important topic for consideration concerns the broadening of the instructional curriculum accompanying the use of three tests for state assessment. A finding from the Third International Mathematics and Science Study (TIMSS) may be relevant here. Nations that excelled in mathematics in the TIMSS assessments have a more narrow but focused curriculum than the United States. The U.S. curriculum appears to be unnecessarily broad, but shallow.<sup>22</sup> Arizona's teachers may face a broader but shallower curriculum in 2005 because they have three websites to visit for guidance on test preparation for their students. Given that the National Assessment of Educational Progress (NAEP) and Arizona Instrument to Measure Standards (AIMS) are mandated, the goal to better focus instruction may be served by the elimination of the TerraNova (TN) exam. Although instruction focused on a narrower curriculum may result in higher test scores, there is no evidence that overall student learning is improved by this focus.

National standardized norm-referenced tests (NRTs) are intended to provide information for a state about its students'/schools'/districts' performance relative to a national norm. However, because the validity of the information NRTs generate is in

question, its usefulness is also in question. Because NAEP is considered a more valid source of information about state assessment trends, NRTs that do not show trends similar to NAEP do not serve their intended purpose.

For comparing student performance across subjects, it is necessary to have scores that are comparable across subjects. The Arizona Department of Education (ADE) can provide comparable scores by creating state user norms for the AIMS scale scores and including this information when reporting the levels of performance. Because all schools within Arizona teach the same content standards and have the same information about test preparation, the students in Arizona comprise a well-defined group for creating norms. ADE could use other types of scores for the purpose of reporting comparable scores, but percentile ranks (PRs) are recommended. When norms are developed for this purpose and used by ADE, AIMS' PRs will be the preferred scores for comparing student performance across subjects. PRs included with each student's score will allow meaningful profile interpretation and reduce the misinterpretations that currently exist regarding school and student performance in mathematics, reading, and writing.

TN scores to can be used to compare performance across subjects; however, the full-length TN will only be administered at grades 2 and 9 in 2005. For other grades, only a subset of TN items will be used at any grade level, and the results will be less valid than those obtained with a full-length achievement test. However, if it is desired to use these scores for reporting student achievement profiles, state user norms should be used for that purpose.

To compare the averages of scores and the percentage of students at a specific performance level for a given school with other schools in the state, it is necessary to have norms for school comparisons. ADE will have the results of annual testing for developing those norms each year. All schools will administer the Dual-Purpose Assessment (DPA) at the same time and under the same conditions. The average scale score on AIMS and the percentage of students rated "meets or exceeds standards" for each school will be available soon after AIMS testing is completed each year. These averages and percentages can be compiled into distributions to produce PRs for schools;

these results are known as state user norms for schools. This information could be made available on the ADE website, but more importantly, should be included on each school's status report to enhance interpretability of the school achievement profiles.

Perhaps the diminishing role of NRTs is reflected in the scaling back of Arizona's NRT into a DPA. Because the norms are suspect, the financial cost of NRTs is difficult to justify. The underlying issue is, as Nitko suggested, "whether tests used in any improved accountability scheme will help students learn."<sup>23</sup> Until adequate evidence is available, showing that a heavy concentration on testing actually results in educational improvement and student learning, perhaps Arizona should allocate their resources more judiciously. Given the cost of the NRTs that are not federally mandated, the unintended consequences of high-stakes testing are important issues to consider.<sup>24</sup>

The addition of AIMS in grades 4, 5, and 7 is improvement not only because of the NCLB requirement, but also because testing in continuous grades provides information for a comprehensive database as suggested by Haladyna.<sup>25</sup> Arizona's interests are served by focusing on the testing mandated by No Child Left Behind (NCLB) and creating user norms for AIMS.

The status of Arizona students compared with students in other states is a more complicated issue. The use of NRTs creates a credibility gap; using the TN will not remove the problem that arose when the Stanford 9 was used. Because NAEP is the only test that has credibility with the national press, it makes sense to use NAEP to determine Arizona's standing within the nation. Although the federal funding of the NAEP assessment may be seen as saving the individual states the financial burden of additional assessment, NAEP does not yield the individual student or school information required by NCLB. The limited administration of the NAEP creates a dilemma that cannot be resolved by any state.

## *Recommendations*

It is recommended that:

1. The Arizona Department of Education (ADE) create state user norms for AIMS and the TN each year for grades 2 through 10. These norms can be used to provide a percentile rank for each reported score, an addition that will generate meaningful achievement profiles.
2. ADE clearly specify which measures are used for what purposes. For example, AIMS data are for student and school comparisons across subjects within Arizona, as well as for reporting how students and schools fare regarding the levels of performance. TN provides information for grades and subjects where National Assessment of Educational Progress (NAEP) tests are not available (if TN trends are found to fit NAEP trends), and supports comparisons between subjects from AIMS. NAEP data are for comparing Arizona with other states, and for verification of reasonable standards.
3. ADE compare schools using norms for schools, not with norms for individual scores, when reporting the results of annual testing. Demographic information on students and schools relevant to the interpretation of test results should be included.
4. ADE conduct a study to recommend the number of curricula that teachers are expected to use as guides to instruction. The state content standards are the focus of instruction, but the NAEP objectives cannot be ignored because NAEP is required by federal law. If it is determined that fewer curricula provide better focus for instruction, the Arizona State Legislature should consider legislation to eliminate the requirement for a national standardized test.



15 See [www.nagb.org](http://www.nagb.org)

16 Burstein, L. (1990). Looking behind the “average”: How are states reporting test results? *Educational Measurement: Issues and Practice*, 9, 23-26.

17 *Ibid.*

18 Linn, R.L. (2000). Assessments and accountability. *Educational Researcher*, 29, 4-16.

19 Skinner, R.A. (2005). The state of the states. *Education Week*, 24, 77-137.

20 Education Watch (2004). Educational trust. Author. Retrieved January 12, 2005, from <http://66.43.154.40:8001/projects/edtrust/index.html>

21 Green, J.P., The Manhattan Institute of Policy Research. Retrieved January 12, 2005, from <http://www.manhattan-institute.org/html/greene.htm>

22 R. Gallimore (personal communication, February 24, 2005). Ron Gallimore is a professor at UCLA who studies the effect of curriculum on student achievement on TIMSS.

23 Nitko, A.J. (1990). Lake Wobegon revisited. *Educational Measurement: Issues and Practice*, 9, 2.

24 Nichols, S. & Berliner, D. (2005). *The inevitable corruption of indicators and educators through high-stakes testing*. Tempe, AZ: Education Policy Research Unit, Education Policy Studies Laboratory, Arizona State University. Retrieved April 20, 2005, from <http://www.asu.edu/educ/eps/EPRU/documents/EPSTL-0503-101-EPRU.pdf>

25 Haladyna, T.M. (2004). The condition of assessment of student learning in Arizona: 2004. In A. Molnar (Ed.), *The condition of Pre-K-12 education in Arizona: 2004* (Doc. # EPSTL-0405-102-AEPI). Tempe: AZ: Arizona Education Policy Initiative, Education Policy Studies Laboratory, Arizona State University. Retrieved February 12, 2005, from [http://www.asu.edu/educ/eps/AEPI/AEPI\\_2004\\_annual\\_report.htm](http://www.asu.edu/educ/eps/AEPI/AEPI_2004_annual_report.htm)