Improve student assessment measures

Today’s students are put through an unprecedented and confusing myriad of tests and other assessment measures. While assessment is generally identified as the key purpose for today’s extensive testing systems, the reality is far more complicated, often leaving students, parents, educators, and policymakers confused and frustrated.

Educators know the importance of student assessment, which is why PSEA supports tests that measure what is being taught and that identify where students need additional instruction or assistance. PSEA supports high standards that clearly define what students should know and be able to do.

It’s often said that tests should be used as a stethoscope, not a hammer. The advent of high stakes testing – using only the scores resulting from one test on one particular day, to judge, label or determine important decisions – is destructive and counterproductive. Tests can and should provide useful and valuable information about what students are learning, but it is not helpful to use them to punish or stigmatize students, teachers, or schools.

Greater understanding of the wide range of tests and assessment measures that are mandated and used for Pennsylvania’s students – both federally and by the Commonwealth – will enhance our collective ability to make better choices for students and public education.

Pennsylvania System of School Assessments (PSSAs)

The Pennsylvania System of School Assessments, also known as PSSAs, is a series of tests given to students in grades three through eight and grade 11. The assessments are in math, reading, writing, and science. The PSSA tests are designed to determine what students know and are able to do compared to the Pennsylvania Academic standards for each grade level. The tests are standardized, combining multiple choice and constructed response questions. Nearly a decade after the introduction of the PSSAs, the Pennsylvania Department of Education (PDE) undertook a performance standards-setting process in order to change the reporting of PSSA results from a norm-referenced to a criterion-referenced system. Consequently, instead of reporting scaled scores or percentages scoring in each quartile of the performance distribution, the main reporting of PSSA results was subsequently referenced to four descriptions of performance (with respect to Pennsylvania’s academic standards): Advanced, Proficient, Basic, and Below-Basic. After passage of the Elementary and Secondary Education Act (ESEA), federal legislation commonly known as No Child Left Behind (NCLB), the PSSA tests primarily have been used to determine whether a child is at, above, or below the proficient standard for that grade’s PSSA exam.
The math and reading assessments are generally seen as high-stakes assessments for schools and districts in that the results are used to determine whether the federally-mandated adequate yearly progress (AYP) is made under NCLB. The science and writing assessments are not used to calculate AYP.

AYP is determined by the percent of students scoring Proficient or Advanced as determined by the “cut scores” (e.g., the score separating Proficient from Basic performance) for each assessment.

The PSSA testing program has undergone a steady evolution since its inception in 1992 as a school-level evaluation model. More recent changes, accelerated by the adoption of The Pennsylvania Academic Standards for Reading, Writing, Speaking and Listening and Mathematics in 1999 and passage of the NCLB in 2001 have moved the PSSA increasingly in the direction of individual student assessment.1

PSEA believes that strategic planning and meaningful evaluation are essential to the role of PSSAs in public education. Without them, the use of these standardized tests could actually have a negative impact on teaching and learning in the Commonwealth.

**PSEA Recommendations**

- Require that PDE explicitly state the purpose(s) of all standardized tests and fully validate the tests for each purpose (Content, Criterion, and Consequential validity should all be addressed). This should be done by expert outside researchers at five to 10 year intervals.

- Reset all PSSA cut scores for passing and other categories independent of the existing scores in a manner consistent with the highest standards of the educational measurement profession. These cut scores should also be subject to criterion and content validation.

Given the increasing consequences that have been attached to the PSSA tests and the greater ones that will be attached to the forthcoming Keystone Exams, it is essential that the performance levels on all exams be evaluated, validated, and if indicated, adjusted after comparing test performance data with other measures of educational outcomes such as postsecondary educational achievement or labor market attainment. It would be arbitrary and illogical to maintain cut scores at levels that prove to have little meaning with respect to the ultimate educational outcomes we hope to promote.
High stakes impact of current use of PSSAs

The accountability measures in ESEA/NCLB (public school choice, private management, conversion to charter schools, and dismissal of staff) all are activated based on the failure of Local Educational Agencies (LEAs) to make AYP by moving toward having all students reach the designated proficient level on their respective state’s assessment. Under ESEA/NCLB, states must have at least 95 percent of enrolled students participate in their testing program. This requirement can place a particularly high burden on LEAs that serve unusually high proportions of at-risk or special needs students.

Pennsylvania has, by design, set relatively, if not artificially high performance standards and cut scores needed to reach the proficient level. Most of the cut scores recommended through the initial standards-setting process were arbitrarily raised by a quarter of a standard deviation upon the recommendation of the PDE to the State Board of Education. In practical terms, PDE staff who oversaw the initial development of the PSSA math and reading performance levels indicated they intended that a proficient scorer on the 11th grade exams would be able to undertake college level work without remediation, which is more than a year prior to that point in a student’s education.

It is important that high-stakes policy benchmarks be set at challenging – but realistic – levels. Unfortunately, benchmarks set for AYP overlook important factors, and unintentionally raise the stakes for students who confront the most significant obstacles to learning. Under the ESEA/NCLB regulations, the US Department of Education (USDoE) arbitrarily decided that only one percent of all students (those with the most severe cognitive disabilities) could have
their advanced or proficient scores on alternative assessments (in Pennsylvania, the PSSA) count towards AYP. An additional two percent can have their advanced or proficient scores on a modified assessment, PSSA-M, (same grade level standards) count towards AYP. It is noteworthy that in 2007, 0.6 percent of all 11th graders taking the PSSA were classified as having mental retardation. That does not include the much higher proportion whose IQs are close to that category but must take the PSSAs. One should keep this in mind when considering that testing proponents frequently express the goal to have these tests reflect college readiness which is often and controversially equated with career readiness.

Dr. Ronald Hambleton, an internationally renowned expert on testing and standard-setting who was a member of the Technical Advisory Oversight Committee that provided ongoing advice to both PDE and the State Board of Education (State Board) throughout the initial PSSA performance level setting process in 2001, has offered 20 criteria by which standard setting should be evaluated. His seventh criterion highlights the obvious problem with the adoption of the PSSA proficiency cut score for NCLB or other accountability purposes:

…Were panelists explained the purposes of the educational assessment and the uses of the test scores at the beginning of the standard-setting meeting? …(A briefing on the uses of the assessment scores and the assessment itself and scoring is fundamental for panelists to set appropriate performance standards. Very different standards may result depending on the purpose of the assessment. For example, were the purpose of the assessment principally diagnostic, panelists might be expected to set fairly high standards to maximize the number of examinees who might receive assistance. A very different set of performance standards would result if the same test were being used to award high school diplomas.)

It would have been impossible for the original standards setting advisory committees (Bookmark and Borderline Groups) as well as PDE, or the members of the State Board to have known the NCLB uses of the PSSA cut scores, because the NCLB Act was passed nearly eight months after the State Board adopted these regulations. In fact, PSEA possesses anecdotal evidence that Bookmark panelists understood that Basic achievement would be considered the minimum passing score, which in fact turned out to be the case under Act 16 of 2000, the Education Empowerment Act. The Secretary of Education at the time later confirmed this was the Department’s view.
According to a wide variety of research-based indicators, Pennsylvania’s public schools are among the best in the nation, and student achievement continues to improve. As we digest the studies and scores that illustrate these significant achievements, it is important to continue to examine them in an effort to determine the underpinnings of these successes. As always, our goal should be to invest in and support programs that are working in our schools.

Certainly, the well-trained, dedicated and professional teachers of Pennsylvania have been central to these achievements. In addition to the high quality of our teachers, recent investments in public education programs that have provided critical teaching resources contributed greatly to this success. Continuing and expanding investments in small class size programs, early childhood education, parental engagement (and other programs identified in this document) is essential. It is equally important to maintain the Commonwealth’s commitment to Act 61 of 2008, which created an adequate, equitable, and predictable basic education funding formula for the first time in nearly two decades.

The key problematic result is that these performance level cut scores became the anchors for the subsequent standards setting exercise in 2005 (which is why the contractor Data Recognition Corporation refers to its adjustment process as performance level validation as opposed to standards setting). Interestingly, in all but a couple of instances, the new panelists recommended substantial decreases in the cut scores for the 3rd, 5th, 8th, and 11th grade PSSA math and reading tests. However, under direction from PDE, the score adjustments could not exceed one standard deviation from the original cut point even if that was the recommendation of the panelists. PSEA is concerned that when the new Keystone Exams are created, the performance levels will again be politically constrained by consideration of the percentages of students who currently fall within the performance categories of the 11th grade PSSA tests. Given the lineage of the PSSA and the fact that the Keystone Exams (unlike the PSSAs) are intended as exit exams which students must pass before they can graduate, this would be inappropriate as well as harmful.

Almost as soon as ESEA/NCLB was passed, researchers asserted what is now conventional wisdom; that is: the goal requiring all students to reach the NAEP or equivalently demanding
state assessment proficient level is unattainable. This prediction is now coming true as LEAs approach the Act’s 2014 deadline for 100 percent proficiency and more and more schools are unable to make AYP.

Value-Added Measurement (VAM)

PSEA believes that effective teacher evaluation is important to ensure quality education, but the use of achievement tests to measure the value teachers add to the education of their students is fraught with problems. Studies of this practice, commonly known as value-added measurement, are critical of its effectiveness because current methods simply cannot isolate the influence of teachers, or measure such influence in a valid or reliable fashion. Current value-added methods cannot establish a causal relationship between individual teachers and the changes in their students’ test scores. Moreover, the “growth in achievement of a teacher’s students” is not a direct measure of the behavior of teachers. In fact, given the changes in topics tested across years within the same general subject areas, value-added methods do not measure student academic growth with respect to specific academic content standards. Using algebra scores from one year and geometry from the next would be analogous to measuring a student’s height in one year and their weight in another and asking, how much did the student grow?

The PSSA tests have not been validated for evaluating teacher performance. However, many have proposed using them for this purpose despite the fact that nationally recognized professional and technical standards require validation of tests for all of their intended uses. It also is important to note that due to the design limitations of the PSSA achievement tests, value-added methods cannot be applied to PSSA results at the academic standard level, thus providing little if any information to help teachers make improvements in instruction.

PSEA is concerned that despite all the unresolved methodological concerns surrounding value-added measurement, the measures resulting from their impenetrably complex quantitative calculations will be given undue weight in all evaluation processes. Regrettably, to many individuals, value-added results will simply, but inaccurately appear objective and scientific.

The use of a value-added model for any screening or signaling purpose should be subject to full, independent, peer review. Because small errors in calculation can lead to large consequences, we do not believe that the proprietary status of any aspect of a system used to measure academic performance should preclude outside review of data, models, computational algorithms, results and reporting. The testing and measurement processes must be fully validated for all their intended uses. With such complex and opaque measurement systems, policymakers, students and their families, administrators, educators, and the public already are being asked to take a great deal on simple faith rather than sound practice or research.
Students with disabilities

Due to a lack of appropriate accommodations, special education students are often prevented from demonstrating what they know when taking the PSSA, which leads to the inappropriate identification of school entities for sanctions under NCLB.

Because it is impossible for the state to identify every permissible and appropriate accommodation, the PDE Accommodations Guidelines must be revised to authorize the use of non-standard accommodations which do not compromise the validity of the test. State officials who are experts in the use of test accommodations must also provide direct and on-going training to those responsible for administering the test, particularly to those responsible for designing or providing accommodation to students with Specific Learning Disabilities or Attention Deficit Hyperactivity Disorder (ADHD).

PSEA believes that the Individualized Educational Plan (IEP) team should dictate which required state and local assessments are appropriate for students with exceptionalities and allow for exclusions and/or alternative forms of assessment.

Keystone Exams – a new test on the horizon

Recently, the State Board enacted regulations establishing a new set of state standardized exams to be known as Keystone Exams.

Beginning with the class of 2014-15, students will have to demonstrate proficiency in the following subjects by passing each course with either a validated local assessment (final exam) or a Keystone Exam as the final exam that is worth one-third of the course grade for: English composition and literature; algebra I; and biology.

Beginning with the class of 2016-17, students will also have to show proficiency in history, civics and government via “validated” local assessments or Keystone Exams. Students in districts that use Keystone Exams will have to pass the following courses, with each Keystone Exam counting for at least one-third of the final course grade:

- English composition and literature;
- Two of the following math subjects -- algebra I, geometry, algebra II;
- One of the following sciences – biology, chemistry; and
- One of the following social studies – American history, civics and government, world history.
These requirements will apply to students in public school districts, charter schools, Area Vocational Technical Schools, and Career and Technical Centers.

PDE will ask the U.S. Department of Education to approve the algebra I, English composition and literature, and biology Keystone Exams as replacements for the 11th grade PSSA. If approved, these three Keystone Exams will be mandated and the 11th grade PSSA will be discontinued. Students will take each Keystone Exam at the end of each respective course in the grade they take the course (that is, they would not all be taken in 11th grade.)

Although PSEA did not support the final regulatory package that established this system of Keystone Exams and “validated” local assessments, we did not oppose it. PSEA worked closely with the State Board of Education and PDE to extensively revise the original proposal that called for 10 paper and pencil exit exams – students would have been denied high school diplomas if they did not score proficient on at least six of the 10 tests.

A report released in August by the Center on Education Policy shows Pennsylvania students were the only students in the nation to make gains in all academic categories from 2002-2008.

Implementation of Keystone Exams

Pennsylvania’s education leaders now need to turn attention to ensuring that the new regulations are implemented effectively and fairly. PDE will be establishing three statewide committees to work on different aspects of the new regulations: the Advisory Committee for the development of performance level descriptors and cut scores; the State Assessment Validation Advisory Committee; and the Local Assessment Validation Advisory Committee.

Each of these Committees will have some public representation. Legitimate input from a broad cross-section of the public must be obtained, in order to ensure broad support to effectively implement that work product.

Perhaps the single most important task to be undertaken by a committee will be consequential validation to be performed on the statewide assessments. That is, validity studies of the Keystone Exams will be performed at least every five years to determine the degree to which the Keystone Exams and performance level cut scores are valid for the purposes for which they are used, and predict college and career success. Pennsylvania and other states have gone quite far down the road of reliance on standardized, paper-and-pencil tests, without checking to see whether a student’s score category on such tests actually is a valid, precise, or reliable predictor of whether he or she does well in college and career. That data will now have to be collected and
studied; if we do not determine whether these test score categories are accurate predictors, students are vulnerable to unfair stigmatization about their preparation for college and career. That has ramifications for students and for their schools.

PDE and the State Assessment Validation Committee will investigate and make a recommendation on the use of a certificate based on industry approved standards and performance on a National Occupational Competency Testing Institute (NOCTI) as an alternative pathway to graduation. This recommendation is due to the State Board of Education within one year of the effective date of the regulation. It is critically important to Pennsylvania’s career and technical education students that this report be full and fair and that, if it is warranted, the NOCTIs be added as an alternative way for career and technical education students who have used their high school careers to prepare themselves for their post-graduation careers to meet graduation requirements.

The state has committed to pay one-half the costs of local assessment validation. Each school district will have to pay the other half of these costs. We all have to work to make sure the local assessment validation process is a productive, useful process that results in improved local assessments in places where improvement is needed. The end goal should be to produce richer, more effective local assessments so that students benefit from the new regulations.

At the same time assessment work is proceeding, the very best contribution Pennsylvania’s leaders can make is to keep attention on funding and on implementing education initiatives that are proven to improve student achievement. Educators will be eager to work with policy leaders to implement these programs.
Multiple measures

Student performance is much more than a test score on a particular day under a particular set of circumstances. It’s a simple fact of life that human performance is multi-dimensional and varies over time. We would think it absurd to judge sports teams on the basis of a single game rather than on team performance over the course of a season. It doesn’t make any more sense to determine the fate of a student, or the livelihood of a public school professional employee, based on a snapshot from a single test.

We think there is a better way. There are a variety of measures that, in concert with test scores, would provide a much more accurate picture of school performance, including: graduation rates of at-risk students; Advanced Placement (AP) participation rates and AP scores; the percentage of students continuing their education and training beyond high school; achievement of goals set by school employees, whether these pertain to reduction of dropout rates, the successful completion of a portfolio of course work or a culminating course project; the successful completion of parental/community involvement programs; and/or the successful implementation of a new curriculum. The state should provide a comprehensive list of possible indicators of performance, but the schools should focus on those indicators that are most relevant to their goals.