COMMENTS BY THE
ASSOCIATION OF TEST PUBLISHERS

In Response to an Informal Request Posted on
Homeroom, the Official Blog of the US Department of Education
"Help ED Improve How We Evaluate State Assessment Systems"

August 7, 2013

The Association of Test Publishers ("ATP") files these comments in response to the Official Blog of the United States Department of Education ("Department"), as noted above. The Department is seeking comments, especially from assessment experts, on how it reviews and approves state tests in light of the move to a new generation of tests linked to college- and career-ready standards, including the Common Core State Standards ("CCSS"), including tests being developed by the two Race-to-the-Top Assessment consortia, as well as individual state tests that are not part of the consortia. These comments are submitted timely by the due date of September 30, 2013.

The ATP is the international trade association representing some 175 publishers and developers of assessments (both non-profit and for profit) used in a variety of settings, including virtually every educational purpose for which the Department of Education is responsible. The ATP has served as the "Intelligent Voice for Testing," providing input to the United States Congress, state legislatures, and federal and state agencies in their efforts to examine issues surrounding testing and the use of test data. Many ATP members provide assessment products and services under contract to states and the assessment consortia and are then called upon to provide evidence used in the peer review process.

These comments are submitted on behalf of educational assessment members, including virtually all of the entities that provide testing products and services to the states under the No Child Left Behind Act and related laws and regulations administered by the Department. The members of ATP have supported the Race-to-the-Top ("RTTT") assessment initiative and the alternate assessments based on alternative achievement standards funded by the General Supervision Enhancement Grant program by providing vital services to the various consortia developing assessments.

The members of the ATP have many decades of experience in developing and implementing complex assessment systems in all 50 states and the nation's 15,000 school districts. Those testing companies work closely with their SEA clients and the consortia to ensure that statewide testing programs are implemented and operated in accordance with all federal and state regulations. The U.S. testing industry comprises educators, researchers, psychometricians, and technologists with extensive experience in developing and administering technically sound assessments that are used for many different purposes.

Inasmuch as the Department has currently suspended use of Peer Review, recently issued guidance to the consortia on technical review of Race to the Top Assessments, and has defined "high quality assessments," the ATP submits that it is most appropriate to synthesize these approaches in order to articulate a single clear and consistent set of criteria. The ATP hereby submits these comments to suggest ways of improving the peer review process to ensure that each statewide assessment enables valid and reliable inferences to be drawn about student performance so that the assessment system meets the requirements of the ESEA and to ensure that the peer review process is fair, provides due process, and is not burdensome on states. Furthermore, the process should be guided by principles of transparency, and should apply the same criteria for general and alternate assessments, and for single state assessments and consortia assessments. We strongly feel that our comments would improve the peer review process to ensure that state assessment programs benefit students, teachers, administrators, and all other stakeholders in determining the effectiveness of our educational system and documenting the student achievement.

The Department, in its Homeroom blog, raises several important questions to which we will respond:

- Are there models or best practices in conducting peer reviews that are applicable and practical for state assessment systems?
- What types of evidence can and should a state provide to demonstrate that its system meets the elements of a high-quality assessment system?
- What benchmarks or rubrics can the Education Department establish to help evaluate the evidence submitted by states?
- Are there components of the department's current process that can or should be revised or are there aspects the department should add?

<u>Discussion of Best Practices</u>

The Department has requested input on changes to criteria for conducting peer review of statewide assessments now that Race to the Top Assessments and alternate assessments being developed by various consortia will be available for state use in the 2014-2015 school year. It is not surprising that many states will be making substantial changes to their accountability systems to reflect use of the CCSS, as well as the assessments being developed to measure them. We believe that the peer review process for evaluating these new assessment systems should be consistent with the overall intent for implementing the new standards, thus providing consistency of effect. We also note that science assessments are now part of state assessment programs that will also require peer review.

The Spring 2014 field-testing of Smarter Balanced and PARCC assessments will showcase new item types, many of which are technology-enhanced items (TEIs), aligned to the CCSS. The use of new items, in particular TEIs, are designed to make assessment more engaging for students and to measure deeper levels of constructs required by the CCSS with greater efficiency. In many cases, TEIs that can be computer scored can replace constructed response items that require hand-scoring. TEIs lend themselves to instruction, as well as to assessment. It is important to note that these technologies are not only present on the PARCC and Smarter Balanced assessments, but are also used on an increasingly wider basis by states in their state-specific assessments being built by our members. We submit that the Department must evaluate on a consistent and objective basis the technical quality of each state assessment that incorporates such new items in order to ensure their reliability and validity, regardless of whether a state is a member of an assessment consortium. In that sense, the long-standing requirements under Title 1 of the Elementary and Secondary Education Act (ESEA), for states to ensure that their assessments are valid and reliable, remain in full force and effect.

To implement the Improving America's Schools Act, which established the core requirements for statewide assessments under the Elementary and Secondary Education Act, the Department adopted the peer review process nearly twenty years ago as the means to evaluate state assessment systems. Nothing about the RTTT Assessment program changes these core statutory requirements for state assessments. As the Department has recognized, peer review applies equally to each state assessment program, regardless of what assessments are part of that program – and who developed those assessments. Consequently, the peer review process must be transparent, fair, and free from all bias and subjective decisions, adhering to objective and technically-sound principles.

It is unclear if the Department intends to place any reliance on the newly issued guidance given to the Consortia grantees for how the Department will evaluate the consortia assessments. *See* "Race to the Top Assessment Technical Review" (April 2013). The ATP applauds this flexibility in presenting technical evidence (Part II, describing how "examples could include" evidence) and the higher degree of focus on Test Security issues (Part II (3)). Indeed, the recently updated *Operational Best Practices for Large Scale Statewide Assessment Programs* (September, 2013) (jointly sponsored with the Council of Chief State School Officers), includes significant expanded materials on test security (*id.* at Chapter 8). As useful as these review elements may be, they must apply equally to both consortia assessments and non-consortia assessments. For example, in its April 2013 guidance document the Department observes that the consortia assessments will not be able to provide evidence on DIF and on alignment (page 2) and states that such evidence will not be required at an early stage of review. This kind of flexibility makes sense in a time when there are major changes to the standards and therefore to the assessments. We believe that the flexibility should not be restricted to the consortia's assessments, but should be extended to all new assessment systems, including those being developed or adopted by individual states. Finally, the ATP notes that the Appendix chart for use by the reviewers is virtually identical to the current peer review chart. As such, there is a potential disparity between "recommendations" and "summary" that could be used to overlook an overall approval for the sake of focusing on specific issues highlighted in the recommendations. We believe that continuing this type of review process can politicize the review process and therefore the Appendix should be modified, as discussed below.

The ATP cautions that the peer review process must not be used to advance any political or ideological purposes – the Department must assure that peer reviewers do not use the review process as a means to advocate for Consortia assessments in preference to assessments built by commercial publishers under contract to individual states. For example, the Race to the Top Assessment Technical Review comments that "both the USED and the consortia have a vested interest in the success [of theses assessments]…." We reiterate that all state assessment systems should be reviewed on the same level playing field regardless of whether they are implementing the Consortia assessments or assessments they have developed. As a matter of fact, one could argue that the Department's strong public support of PARCC and Smarter Balanced has been used as the basis for attacks from those who see it as a federal intrusion on states' rights. Maintaining a fair and consistent process for peer review of all assessments, regardless of source, would help to support the state-led spirit of the entire CCSS movement.

Regarding the comment in the blog that, "These new systems are in direct response to educators and parents asking for assessments that are more than just 'bubble tests,' and provide better information to inform and improve teaching and learning in our classrooms," we submit that the majority of items in development by the consortia will continue to be similar to the traditional items currently seen in state accountability tests. Accordingly, there should be no reason to prefer a consortium assessment over currently available statewide assessments merely because they contain TEIs. This comment overlooks the reality of state assessment systems that have multiple item types and that well-developed multiple-choice items are capable of assessing depth of knowledge.

Heavy emphasis is being placed on the new assessments aligned to the CCSS, most notably the goal for all students to be college and career ready upon graduating from high school. It is imperative that the peer review process be as rigorous and comprehensive as possible. When evaluating the new accountability tests, the peer review process should consider an item's operational or field-test status, as well as other relevant factors. Typically, a test has a singular purpose that is described fully prior to development of the test. While a test may have dual or multiple purposes and be validated for them, the Department should be cautious in approving assessment systems that attempt to utilize a single measure for multiple purposes, because the validity of the test results will be diminished by overextension of the assessment, perhaps even for purposes for which the assessment has not been validated. Best practices that can guide peer reviewers in understanding test design, item development, item banking, form development, and other related factors, especially for technology-based assessments, are found in the CCSSO/ATP *Operational Best Practices.* We encourage the Department to use these best practices in training peer reviewers.

Proposed Changes to Evidence and Benchmarks

The current review process has a number of categories by which a test or assessment system should be evaluated. These seven broad categories should serve as a starting point for the next generation of peer review. The seven areas or categories comprise the following:

1. Content Standards
2. Academic Achievement Standards and Standard Setting
3. Statewide Assessment System
4. Technical Quality
5. Alignment
6. Inclusion
7. Reporting

Each of these seven categories contains a number of specific criteria, although it might be more efficient to use fewer criteria that are more principled than prescriptive in nature.

Consequently, we have suggested some modifications to the criteria, as set forth below, against which peer review panels would evaluate each state assessment system.

1. Content Standards
   a. An inclusive and explicit process consistent with all applicable state requirements for adopting the content standards was followed.
   b. All major stakeholders were involved in the development and/or adoption of the content standards.
   c. A determined, comprehensive, and well-documented plan was implemented to disseminate the content standards.
   d. A thorough and well-documented plan was implemented to ensure that all students had comparable access to instructional materials aligned to the state content standards and professional development and resources were available to help teach the new content standards so that all students have a fair and adequate opportunity to learn prior to assessment.

2. Academic Achievement Standards and Standard Setting
   a. All major stakeholders were involved in the development and adoption of the academic achievement level standards and achievement level descriptors (ALDs).
   b. The standard setting method used to establish cut scores for the ALDs was well documented and implemented.
   c. Standard setting was carried out using content area experts from as many of the major stakeholder groups as was practical (e.g., teachers, administrators, specialists from special education, and English language learner (ELL) teachers).
   d. The process for adopting the resultant cut scores was well-documented and followed.

3. Statewide Assessment System
   a. A rational, comprehensive system of assessment at the state and local levels was designed and implemented that allows all students to demonstrate their learning of the content taught.
   b. Special attention was given to student access to the system, and reasonable accommodations were put into place to assess all students.
   c. All stakeholders in the state were part of the design and implementation processes of the assessment system.
   d. Local assessments used in accountability meet the same requirements as the state-level assessments and contribute new and different information that help teachers and students understand strengths and areas in need of improvement.

4. Technical Quality
   a. The assessment under evaluation followed the *Standards for Educational and Psychological Testing* and the CCSSO/ATP *Operational Best Practices (2013),* including a foundation of internal consistency/internal validity evidence is provided through technical documentation based on the joint standards and necessary internal validity arguments/reliability.

   b. The psychometric quality of the items and the test as whole supported the purpose of the statewide assessment system.
   c. Both item and test reliability was demonstrated through standard item and test analyses as appropriate for each assessment (e.g., point-biserial, Chronbach's alpha, and classification consistency).
   d. Differential Item Functioning (DIF) was provided, when possible, for all relevant groups, with at least 50 students belonging to the focal group.
   e. Validity was demonstrated in terms of both content validity of the test as well as validity arguments for each specified use of the assessment results (e.g., accountability for school and districts, evaluating student growth, evaluating course credit, and evaluating teacher effectiveness).
   f. Linkage to college and career readiness was demonstrated through ongoing research into the success of students exiting public high schools.
   g. Accommodations were accompanied by research, as appropriate, to address the effectiveness of the accommodations and the validity of results when they are used.
   h. Credible evidence is provided that scores are reasonable for the assessment by a strong focus on external validity evaluation of their testing program in terms of student outcomes (e.g., correlations/common patterns between outcomes of state tests to nationally normed tests, NAEP, or international tests, evidence of student outcomes in other content areas).
   i. Evidence of non-arbitrary relationship between society's values for education and test results (e.g., content reviews, standard setting).

5. Alignment
   a. The alignment of the test and test items to the content was documented using a well-documented methodology (e.g., expert review of content alignment relationships between content of tests and Common State Standards/international standards).
   b. The test covered the breadth of the content standards as well as the depth of student learning.
   c. Evidence is provided of expert review of standard setting process and results in line with stated performance standards (PLDs), with less emphasis on the processes of developing the tests beyond necessary industry-standard technical documentation and more focus on outcome comparisons.)
   d. Major stakeholders were part of the alignment studies.
6. Inclusion
   a. Evidence is provided to show that accessibility was a consideration during test design and item development.
   b. Evidence is provided on student participation in the assessment.
   c. Reasonable accommodations were made so that all students could be assessed.
   d. Accommodations were used that addressed individual students' access needs and that did not fundamentally change the nature of the construct being measured by the assessment.

7. Reporting
    a. Individual student reports were provided that provide information consistent with the purpose of the test and that can be supported by the test's psychometric characteristics.
    b. Teacher- and school-level reports were provided that aggregated and disaggregated student results to help inform teachers and administrators as to what is being taught well and what areas are in need of improvement.
    c. State-level aggregate and disaggregate data for all students in the state were provided to give an overall picture of how the educational system is functioning.
    d. A benchmark report was provided that anchored the state assessment results to those of the National Assessment of Educational Progress (NAEP) at grades 4 and 8 in Reading and Mathematics. International benchmarking might also have been presented, linking state assessment results with that of the global community.
    e. Benchmark results in terms of college readiness were provided, showing student results in comparison to academic levels needed for success in college.
    f. Benchmark results in terms of career readiness were provided, showing student results in comparison to academic levels needed for success in various career clusters or pathways.

The ATP offers these recommendations, which we feel are fully consistent with the Department's current definition of "high quality assessments."

Proposed Changes to the Peer Review Process

As pointed out above, the peer review process must be transparent and objective (i.e., unbiased and devoid of politics), and provide comprehensive analyses of state assessment systems. In order to achieve these goals, we submit that each peer review panel should comprise two psychometricians, two academic content specialists, one educator from the special education field, one educator specializing in ELL instruction, one assessment technology expert, and a chairperson who could come from any of these categories or from administration. Each panelist must be vetted to ensure that the individual can conduct an assessment evaluation without any appearance of a conflict of interest and that each possesses the requisite expertise and qualifications; training should be available to all panelists to ensure that these objectives are met. This composition, vetting, and training would ensure that evidence submitted by a state is reviewed consistently between panels, by truly objective reviewers who possess all of the relevant expertise and skill sets needed to make a full, accurate, and impartial evaluation of each assessment system.

The peer review panel should make a single overall recommendation for the assessment system (i.e., Approved, Approved with Recommendations, or Not Approved). The panel report to the Department should contain a consensus overall summary of the state assessment system and a short summary for each of the seven criteria. Beyond this summary, the peer review panel should make recommendations for each of the seven criteria. Therefore, a state may receive an overall Approved with Recommendations in which some of the criteria have been approved while others have recommendations for improvement. This information should be shared directly with the state personnel associated with the assessment system. The Department would better serve a state in understanding exactly what the panel has found by having in-person meetings when sharing this information, at least when a "non-approved" rating is involved.

Next, the peer review process needs to be consistent regardless of whether states belong to a consortium or not. Even though they are using the same tests, states belonging to consortia may have slightly different administration requirements, accommodations, or other system differences that might affect peer review. While the review of assessment systems for all of the PARCC states, for example, could be streamlined with reference to issues about coverage of the CCSS and technical quality of the actual tests, all states should need to submit their assessment system to the Department for approval. Given the nature of the assessment systems that will be available over the next few years, it is likely that state assessment systems will never be fully approved and that one or more criteria areas are works-in-progress. The most obvious criteria where this is true would be "technical quality." It could be argued that the validity of the use of an assessment system is an ongoing process. In this regard, the ATP recommends that the Department establish an independent group to select and oversee the peer review panels and to help insure the consistency of the peer reviews. This group could be set up as a National Technical Advisory Review Panel, appointed by the Secretary in consultation with states, representatives of the psychometric and measurement communities, and the testing associations. The makeup of such a group would be similar to that of the peer review panels and would interact directly with the Department to ensure that the goals and objectives of the revised peer review process are carried out.

Finally, the peer review process must provide appropriate and adequate due process for a state to challenge the decision reached by the panel. Such due process should include the right to petition for reconsideration of a peer review decision, as well as setting forth a process by which a state is entitled to seek judicial review of a non-approval decision.

Conclusion

The ATP submits that many specifics about the Peer Review process still need to be developed and implemented. This document provides only some overarching suggestions that we believe could help in the redesign and implementation of the peer review process. We, and our members, stand ready to assist the Department in creating a reliable, valid, and transparent review process that is consistent, predictable, that is not burdensome on states, and that will benefit students, teachers, administrators, and all other stakeholders in determining the effectiveness of our educational system and documenting the achievement of our students. The future rides on getting this part of the educational system right.

These comments are respectfully submitted by the Association of Test Publishers on September 30, 2013.

ASSOCIATION OF TEST PUBLISHERS

William G. Harris/AST

William G. Harris, Ph.D.
CEO