We are providing these comments in response to the U.S. Department of Education’s August 7, 2013 request to help the Department improve how it evaluates state assessment systems. The mission of the Center for Educational Assessment at the University of Massachusetts Amherst is to improve educational assessment practices and help promote valid assessment of individuals, and valid use of test scores by educators, policy makers, and other stakeholders. We believe the Federal peer review process is a key factor in promoting quality assessment practices at the state and national levels, and we are pleased to have the opportunity to provide feedback as the Department considers improvements to the peer review process. The faculty and staff at the Center have worked with educators and policy makers at the local, state, national, and international levels for several decades. Our comments are based on our experience in working with these colleagues, as well as over 40 years of research on educational assessment.

Our comments are organized into three categories: (a) suggestions for developing a new framework for reviewing statewide (and consortia-wide) assessment practices, (b) reactions to the current process of reviewing statewide and consortium-based assessment systems, and (c) suggestions for improving the feedback the review process will provide to the states and consortia it reviews. The suggested framework for reviewing statewide assessment systems is based on the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999).

A Framework for Reviewing Educational Assessment Systems

A review of statewide and consortium-based assessment systems goes beyond traditional test validation. In addition to evaluating whether test scores are valid for their intended purposes, a review of an educational assessment system involves evaluating whether the goals of the system are being realized. Some may describe this as evaluating the system’s “Theory of Action.” We believe the Federal review system should involve a review of the assessments and
a review of the goals of the assessment program that extend beyond the qualities of the assessments and their scores.

For over 50 years, three organizations—AERA, APA, and NCME, have put forward standards and guidelines to promote fair and appropriate test development practices, and to provide guidance for evaluating tests. The current version of the Standards (AERA et al., 1999) defines validity as, “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (p. 9). This version also describes the process of validation (test evaluation) as developing a comprehensive argument that test scores are appropriate for their intended purposes. This argument-based approach to validation is evident in the following excerpt:

A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses…Ultimately, the validity of an intended interpretation…relies on all the available evidence relevant to the technical quality of a testing system. This includes evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all examinees… (AERA et al., 1999, p. 17)

The Standards also provide a framework for validation by organizing validity evidence to support the use of a test for a particular purpose into five sources: validity evidence based on (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) testing consequences. We recommend the Federal review process adopt this consensus-based framework. Although the new version of the Standards is due out in 2014, it has been confirmed the definition of validity and the five sources of validity evidence will remain.

Basing the Federal review process on the Standards’ framework adds credibility to the process because the framework is based on a consensus built across three professional associations that studied the process of test validation for over 50 years. To incorporate this framework, the Federal review system would require assessment systems to (a) state their intended purposes, (b) identify potential improper uses of test scores, and (c) identify which sources of validity evidence are most appropriate for evaluating the degree to which the test scores are fulfilling their intended purposes and are not leading to unintended negative consequences. An example of the types of evidence that could be linked with specific testing purposes is presented in Table 1. This example is taken from the research agenda proposed for the Smarter Balanced Assessment Consortium’s summative assessments.

It is important to note that if the Federal Review process adopts the Standards’ framework based on five sources of validity evidence, the current requirements for alignment research would fall under validity evidence based on test content. The current the Standards and Assessments Peer Review Guidance (U.S. Department of Education, 2009) separates alignment from the technical quality chapter where validity is described, which is unnecessary.
Table 1

Example of Validity Framework Based on the AERA et al. Standards
(from proposed research agenda for Smarter Balanced Summative Assessments)

| The purposes of the Smarter Balanced Summative assessments are to provide valid, reliable and fair information about, | Source of Validity Evidence |
| --- | --- | --- | --- | --- |
| | Content | Internal Structure | Relations w/ Ext. Variables | Response Processes | Testing Consequences |
| 1. students’ ELA and Mathematics achievement with respect to those CCSS measured by the ELA and Mathematics summative assessments | ✓ | ✓ | ✓ | ✓ | |
| 2. whether students prior to Grade 11 have demonstrated sufficient academic proficiency in ELA and mathematics to be on track for achieving college readiness | ✓ | ✓ | ✓ | | ✓ |
| 3. whether Grade 11 students have sufficient academic proficiency in ELA and Mathematics to be ready to take credit-bearing college courses | ✓ | ✓ | ✓ | | ✓ |
| 4. students’ annual progress toward college and career readiness in ELA and Mathematics | ✓ | ✓ | ✓ | | ✓ |
| 5. how instruction can be improved at the classroom, school, district, and state level | ✓ | | | | ✓ |
| 6. students’ ELA and Mathematics proficiencies for Federal accountability purposes and potentially for state and local accountability systems | ✓ | ✓ | ✓ | | ✓ |
| 7. students’ achievement in ELA and Mathematics that is equitable for all students and subgroups of students. | ✓ | ✓ | ✓ | ✓ | ✓ |

Note: ✓ indicates types of validity evidence needed to support stated purpose.
In addition to alignment research being placed in the category of validity evidence based on test content, information related to measurement precision (e.g. score reliability estimates, estimates of decision consistency), equating, and dimensionality (e.g., IRT model fit) could be placed in the internal structure category. Alternatively, separate categories could be established for measurement precision and equating. As the last row in Table 1 implies, concerns for test fairness should be based on content-related evidence (e.g., sensitivity review), internal structure evidence (e.g., DIF analyses), relations to other variables (e.g., differential predictive validity), response processes (e.g., think-aloud studies across accommodated and non-accommodated administrations), and testing consequences (e.g., studies on graduation, retention, and access). Thus, this framework does not limit the types of studies states should do to support specific purposes. Rather, it helps states organize evidence into categories that support a comprehensive validity argument, and it will help the Federal reviewers evaluate the evidence and the argument.

Many assessment systems are currently proposing to use students’ test scores to evaluate their teachers. Such purposes also need to be validated and so the revised Federal review system should require validity evidence for aggregate measures such as value-added scores for teachers, or aggregated percentile ranks. We are concerned that such measures are being used prior to validity evidence being established. All five sources of validity evidence are likely relevant to support the use of these derivative scores, particularly evidence based on relations to other variables indicative of effective teaching.

Other explicit purposes of an assessment system that go beyond the technical qualities of an assessment (e.g., provide professional development for teachers, improve teaching and learning) do not fit well into the Standards validation framework and so in the Federal review system they should be separated out, and handled using a program evaluation framework. This part of the review process could involve audits of system processes to confirm intended products were produced, were disseminated appropriately, were used and valued by educators and students, and had a positive effect on student learning. Other examples of “program evaluation” type evidence could include evaluation of usage statistics of online assessment resources, document reviews, comprehensive surveys of stakeholders and those involved in delivering the system, and comprehensive surveys of users of the resources (e.g., teachers, administrators, students, parents).

Reactions to Current Peer Review System

In considering the Peer Review System associated with No Child Left Behind, we see several positive features as well as a few areas that could be improved. In addition to familiarity with the NCLB Peer Review, we have also recently reviewed other systems for auditing testing programs such as the Buros’ review process, the European Federation of Psychologists Association model, the Dutch Committee on Test and Testing model, and the Test Review
System used in Germany. Our reactions to the U.S. Peer Review process are influenced by elements of some of these other systems.

**Strengths of the Current Peer Review Process**

A major strength of the current Federal peer review process is the depth and specificity of its guidance for state participants. In our review of different audit models, we observed that many did not provide explicit guidance or examples about appropriate – or, equally if not more important, inappropriate – evidence for quality or compliance with respect to the standards against which a testing program is evaluated. In the *Standards and Assessments Peer Review Guidance* (US Department of Education, Office of Elementary and Secondary Education, 2009), the provision of concrete examples of “acceptable evidence,” “possible evidence,” and “incomplete evidence” is an important feature that should be carried through to the new peer review process, whatever its form or content. Note that we are not recommending complete or exhaustive specification of “necessary and sufficient” evidence because states should have reasonable flexibility to determine their own best response to the law and peer review standards based on their unique contexts. Nevertheless, states need to be informed about expectations of quality and equitable assessment systems.

A second strength of the current peer review process is its iterative nature, in which states must continue to seek approval for system changes, even once they have achieved Full Approval. This feature is important because it acknowledges the dynamism both of implementing an assessment system and of evaluating its validity. The accountability systems states first submit after their shift to the PARCC and Smarter Balanced consortia will need tweaking and revision as their implementation continues, even if these initial submissions are strong and receive approval. We encourage the Department to support states in engaging with and evaluating their accountability systems more regularly than they currently do – not because they should constantly be changing their systems, but because validation work is never finished.

**Areas to Focus Improvement**

One shortcoming of the current system is its focus on compliance, rather than quality. A focus on compliance may lead to the approval of a system that “checks all the boxes,” but is of substandard quality. The important point is that compliance with ESEA, although critically necessary, should not automatically conclude an accountability system is valid for its intended purposes (emphasis on “purposes” as plural). Minimally, as mentioned earlier, an assessment system needs to demonstrate the assessments are congruent with the AERA et al. *Standards*. Clearly, it is possible that an assessment system could comply with the law yet stand in violation of the *Standards* (current examples include the law’s requirement that educational assessments be used to evaluate student growth and teacher effectiveness, when neither of these uses yet has a solid foundation of validity evidence).

A second shortcoming of the current peer review process is its lack of transparency or consistency about reviewer rating processes. Anecdotally, we are aware that states and reviewers
both have described considerable variability in stringency and priorities from one reviewer or team of reviewers to the next. The training for peer reviewers is also unknown. It would be good to standardize peer review training and evaluate the reliability of the review process. It would be good for the Department to make training materials for peer reviewers publically available. In addition to building trust into the review process, these documents would also likely help states develop their materials for review.

Another recommendation is to better use the measurement community (e.g., NCME) to help develop training and other resources for peer reviewers. The validity of the peer review system is an important precursor to the validity of any rulings the system produces.

Other Assessment Review Systems

While there are no clear best practices that emerge from other models for reviewing tests, there are some practices and characteristics worth noting that may be useful in the redesign of the Federal review model. For example, some systems use a third party to solve disputes among reviewers. Others use numerical criteria for certain characteristics such as norms, reliability, or criterion-related validity. Standardized rating scales aligned with the validation framework described above may help make the process more uniform across states.

Suggestions for the Revised Review Process

In addition to our previous comments, we have a few other suggestions for improving the peer review process. We recommend the Department further consider the important role of feedback both during the peer review process and once a determination has been made related to a state’s assessment program. We also encourage the Department to put more emphasis on the degree to which states adequately provide informative reports and actionable data to stakeholders such as teachers, parents, and the media.

Providing Feedback During the Process

Similar to the ESEA waiver review process, USED should consider the use of organizations that could serve as preliminary review teams to help states prepare for their submission of materials. The organization(s) should be familiar with state assessment programs and the Department’s expectations. A preliminary review will help ensure that submissions include all relevant information for peer reviewers, and that they receive submissions in a standardized format. Uniformity of submissions allows for a fair and thorough review of submissions, and subsequently for more substantive feedback to states upon a final determination.

Post-Review Feedback

Upon completion of the peer review process, USED should ensure states receive substantive technical feedback from the peer review team. Such feedback requires that the peer review team be
balanced in its knowledge of psychometrics, policy, and content. As such, substantive feedback related to best practices for peer review criteria in each of these areas would ensure that the entire process is a useful endeavor for states, rather than only one of compliance. Implementing of all feedback need not always be required for approval, unless the areas of critical weakness are noted. The point is, the outcome of the review process should provide formative feedback in addition to a summative conclusion.

Reviewing Score Reports

In the course of peer review, states must demonstrate how they (a) prioritize the effective communication of test results, and (b) facilitate appropriate data and test use. Effective communication and provision of data are critical aspects of assessment system quality, because they require states to show that intended users of assessment results have access to results in ways that are useful and understandable. Efficient and effective score reporting also recognizes that different users of test data have different needs and varying familiarity with the language and statistics of testing. The new Federal review process should require states to provide clear evidence of the score report development processes, including how stakeholder input was gathered throughout the process. Where analysis tools for assessment data (such as online portals, digital reporting systems, dashboards, etc.) are made available to intended users, such tools should be built with input from end users. There should likewise be a monitoring system in place to identify implementation and use problems as well as best practices, to support and promote effective and efficient use of test data. Clearly, we believe the revised review process should include evaluation of the quality, clarity, and comprehensiveness of reporting of results.

Concluding Remarks

We are pleased to have the opportunity to provide comments on what will surely be one of the most important initiatives for promoting valid educational assessment throughout the United States. We hope these comments are the beginning of a dialogue between the U.S. Department of Education and the Center for Educational Assessment at the University of Massachusetts Amherst. Our suggestions encourage the Department to focus on validation of the uses of educational tests, not on the tests themselves. The ways in which states use assessment results will influence the types of data the Department receives about the quality of education across the U.S. Therefore, they will delimit the types of inferences that can be made at the state and national levels. By framing the peer review on the evidence available for supporting explicit purposes, we will be able to more fully evaluate the degree to which the assessments, and the assessment systems, are fulfilling their purposes.

1 These comments were primarily authored by Stephen G. Sireci, Ph.D.; April Zenisky, Ed.D., Molly Faulkner-Bond, Maria Fernanda Gandara, and Joshua Marland. Other faculty in the UMass Center for Educational Assessment include Ronald Hambleton, Lisa Keller, Jennifer Randall, and Craig Wells.