*Briefing Paper Prepared for Members of*
*The Congress of*
*The United States*

# Refocusing Accountability: Using Local Performance Assessments to Enhance Teaching and Learning for Higher Order Skills

George H. Wood
Director, The Forum for Education and Democracy
Principal, Federal Hocking High School, Stewart, Ohio

Linda Darling-Hammond
Charles E. Ducommun Professor, Stanford University
Co-Director, School Redesign Network

Monty Neill
Co-Director, Fair Test (National Center for Fair & Open Testing)

Pat Roschewski
Director of Statewide Assessment
Nebraska Department of Education

May 16, 2007

For More Information Contact
George Wood, Forum for Education and Democracy
740-448-4941
www.forumforeducation.org

*Executive Summary*

# Refocusing Accountability: Using Local Performance Assessments to Enhance Teaching and Learning for Higher Order Skills

By George Wood, Linda Darling-Hammond, Monty Neill and Pat Roschewski

Performance based assessments, often locally controlled and involving multiple measures of achievement, offer a way to move beyond the limits and negative effects of standardized examinations currently in use for school accountability. While federal legislation calls for "multiple up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding" (NCLB, Sec. 1111, b, 2, I, vi), most assessment tools used for federal reporting focus on lower-level skill that can be measured on standardized mostly multiple-choice tests. High stakes attached to them have led schools to not engage in more challenging and engaging curriculum but to limit school experiences to those that focus on test preparation.

Performance assessments that are locally managed and involve multiple sources of evidence assist students in learning and teachers in teaching for higher order skills. These tools engage students in the demonstration of skills and knowledge through the performance of tasks that provide teachers with an understanding of student achievement and learning needs. Large scale examples involving the use of such performance-based assessments come from states such as Nebraska, Wyoming, Connecticut and New York, as well as nations such as Australia and Singapore. The evidence from research on these and other systems indicate that through using performance assessments schools can focus instruction on higher order skills, provide a more accurate measure of what students know and can do, engage students more deeply in learning, and provide for more timely feedback to teachers, parents, and students in order to monitor and alter instruction.

Research evidence suggests that in order for performance assessment systems to work, governments must make significant investments in both teacher development and the development of performance tasks. However, this investment is often no greater than the cost of standardized measures. More important, it strengthens teacher quality and student learning. Performance assessment systems can be reliable and valid, having both content and predictive validity when appropriately utilized.

Based on the evidence that performance based assessment better meets the federal agenda of teaching for higher-level skills, reauthorization of NCLB should support and encourage state and local education agencies in developing performance assessments. Congress can amend Section 1111 (b)(3) of NCLB with a new paragraph (D) that authorizes and encourages states to move to performance based assessments and multiple measures incorporated into a system combining state and local assessments. Authorization for adequate funding to support this move should be included in the legislation.

# Refocusing Accountability:
# Using Local Performance Assessments to Enhance Teaching and Learning for Higher Order Skills

Over the past decade, educators, policymakers, and the public have begun to forge a consensus that our public schools must focus on better preparing all children for the demands of citizenship in the 21st century. This has resulted in states developing 'standards-based' educational systems and assessing the success of districts and schools in meeting these standards measured through more systematic testing. However, most of these tests are multiple choice, standardized measures of achievement, which have had a number of unintended consequences, including: narrowing of the academic curriculum and experiences of students (especially in schools serving our most school-dependent children); a focus on recognizing right answers to lower-level questions rather than on developing higher-order thinking, reasoning, and performance skills; and growing dissatisfaction among parents and educators with the school experience. The sharp differences between the forms of testing used in the United States and the assessments used in other higher-achieving countries also suggest that low international rankings may be related to over-reliance on standardized testing in the U.S.

These unfortunate consequences have occurred despite language in NCLB calling for "multiple up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding" (NCLB, Sec. 1111, b, I, vi). Changing what counts as assessment evidence, coupled with other significant changes in NCLB's accountability structure (e.g., adequate yearly progress and sanctions), could help to overcome these problems and contribute toward school improvement

## Performance Assessment: A Definition

Almost every adult in the United States has experienced at least one performance assessment: the driving test that places new drivers into an automobile with a DMV official for a spin around the block and a demonstration of a set of driving maneuvers, including, in some parts of the country, the dreaded parallel parking technique. Few of us would be comfortable handing out licenses to people who have only passed the multiple-choice written test also required by the DMV. We understand the value of this performance assessment as a real-world test of whether a person can actually handle a car on the road. Not only does the test tell us some important things about potential drivers' skills, we also know that preparing for the test helps improve those skills as potential drivers practice to get better. The test sets a standard toward which everyone must work. Without it, we'd have little assurance about what people can actually *do* with what they know about cars and road rules, and little leverage to improve actual driving abilities.

Performance assessments in education are very similar. They are tools that allow teachers to gather information about what students can actually do with what they are learning – science experiments that students design, carry out, analyze, and write up; computer programs that students create and test out; research inquiries that they pursue, seeking and assembling evidence about a question, and presenting in written and oral

form. Whether the skill or standard being measured is writing, speaking, scientific or mathematical literacy, or knowledge of history and social science research, students actually perform tasks involving these skills and the teacher observes, gathers information about, and scores the performance based upon a set of pre-determined criteria. As in our driving test example, these assessments typically consist of three parts; a task, a scoring guide or rubric, and a set of administration guidelines. The development, administration, and scoring of these tasks requires teacher development to insure quality and consistency. The research suggests that such assessments are better tools for showing the extent to which students have developed higher order thinking skills, such as the abilities to analyze, synthesize, and evaluate information. They lead to more student engagement in learning and stronger performance on the kinds of authentic tasks that better resemble what they will need to do in the world outside of school. They also provide richer feedback to teachers, leading to improved learning outcomes for students.

Extensive research and experience, both here and abroad, have demonstrated that the use of *performance assessments* which are *locally administered* and use *multiple* sources of evidence offer the opportunity to turn assessment systems to serve their primary purpose—**assisting students in learning and teachers in teaching for higher order intellectual skills**. In fact, the assessment systems of most of the highest-achieving nations in the world are a combination of centralized assessments that use mostly open-ended and essay questions and local assessments given by teachers which are factored into the final examination scores. These local assessments--which include research papers, applied science experiments, presentations of various kinds, and projects and products that students construct--are mapped to the syllabus and the standards for the subject and are selected because they represent critical skills, topics, and concepts. Central authorities often determine curricular areas and skills to assess, but the assessments are generally designed, administered, and scored locally.

The *local management* of such assessments refers to both their use and scoring. While not all performance assessments are locally developed many are; and decisions about when to use them in the learning process and how to adapt them to particular content are made at the school or classroom level. This is vital as assessment must be responsive to emerging student needs and enable fast and specific teacher response, something that standardized examinations with long lapses between administration and results cannot do. In addition, as teachers use and evaluate these tasks, they become more knowledgeable about the standards and how to teach to them and about what their students' learning needs are. The process improves their teaching. These rich assessment tasks can also be utilized as formative or benchmark assessments, which help teachers' gauge ongoing progress, while avoiding the reduction of such assessments to commercially available multiple-choice formats.

Using *multiple sources of evidence* refers to the way in which performance assessments provide multiple ways to view student learning. For example, multiple samples of actual writing taken over time can best reveal to a teacher the progress a student is making in the development of composition skills. This provides ongoing feedback to learners as well, as they see how they are developing as writers and what

they have yet to master. In addition, different kinds of writing tasks – persuasive essays, research papers, journalistic reports, responses to literature – encourage students to develop the full range of their writing and thinking skills in ways that writing a five-paragraph essay over and over again do not.

These features of performance, local administration, and multiple sources of evidence are used in many assessment systems. Let's think back to the state driver's license exam. This involves both a written test and a performance assessment on the road. Everyone knows precisely what to expect in terms of the skills to be demonstrated —for example, whether or not the applicant can parallel park—as the examination is not a total secret. The fact that the assessment is open and transparent is not a problem, because the point is to see whether drivers have developed these real-world abilities. The performance is scored by the instructor, working from a rubric, and if the driver is sufficiently successful in all aspects of the examination (as determined by a state cut-off score), a license is conferred. The task is so well defined that instructional programs (driver's education) which include both hands on and classroom instruction clearly demonstrate their effectiveness in preparing students to perform. (This is reflected in the reduced insurance rates we grant to graduates of driver's education programs.) Imagine what life on our roads would be like if we did not require prospective drivers to demonstrate what they know before taking the wheel.

Some states, districts, and schools have constructed a similarly rich set of assessments of competence that measure the higher-order thinking called for by new standards. In many cases they are explicitly intended to augment and complement more traditional tests.

Illinois' assessments provide a good example of the contrast between classroom performance assessment and a state multiple-choice test. The state's grade 8 science learning standard 11B reads: "Technological design: Assess given test results on a prototype; analyze data and rebuild and retest prototype as necessary." The multiple choice example on the state test simply asks what "Josh" should do if his first prototype sinks, with the wanted answer "Change the design and retest his boat." The classroom assessment, however says: "Given some clay, a drinking straw, and paper, design a sailboat that will sail across a small body of water. Students can test and retest their designs." In the course of this activity, students can explore significant physics questions such as displacement in order to understand why what was a ball of clay can be made to float. Such activities combine hands-on inquiry with reasoning skills, have visible real-world applications, are more engaging, and enable deeper learning. They also enable the teacher to assess student learning along multiple dimensions, including the ability to frame a problem, develop hypotheses, reflect on outcomes and make reasoned and effective changes, demonstrate scientific understanding, use scientific terminology and facts, persist in problems solving, and organize information, as well as develop sound concepts regarding the scientific principles in use.

Many states – including Connecticut, New York, and Vermont -- have developed and use such hands-on assessments as part of their state testing systems. Indeed, the

National Science Foundation provided millions of dollars for states to develop such hands-on science and math assessments as part of its Systemic Science Initiative in the 1990s, and prototypes exist all over the country.

Perhaps the most important benefit to utilizing performance assessments is that they assist in learning and teaching. They are *formative* in that they provide teachers and students with the feedback they need from authentic tasks to see if they have actually mastered content. They can also be *summative* in that they can serve as a final assessment of student capabilities with respect to state and local standards. Because of their numerous positive features, they are more sensitive to instruction and more useful for teaching than standardized examinations, while providing richer evidence of student learning that can be used by those outside the classroom or school.

Performance Assessment: Large Scale Examples

As we have noted, it is possible to create and implement assessment systems that include multiple sources of evidence which are performance based and locally managed. Some U.S. states and many countries have developed extensive performance-based assessment systems that engage teachers, parents, and students in thinking carefully about what students have learned and how to measure that learning. Examples include:

- Nebraska utilizes a system of assessments created and scored by local educators. These systems are peer-reviewed in a system supported by assessment experts and include a check on the validity of such assessments through the use of a state-wide writing examination and the administration of one norm-referenced test.
- Wyoming uses a "body of evidence" approach that is locally developed in order to determine whether students have mastered standards required for graduation.
- Connecticut uses rich science tasks as part of its statewide assessment system. For example, students design and conduct science experiments on specific topics, analyze the data, and report their results to prove their ability to engage in science reasoning. They also critique experiments and evaluate the soundness of findings.
- Maine, Vermont, New Hampshire, and Rhode Island have all developed systems that combine a jointly constructed reference exam with locally developed assessments that provide evidence of student work from performance tasks and portfolios.
- In New York, the New York Performance Assessment Consortium is a network of 47 schools in the state that rely upon performance assessments to determine graduation. (Because of the quality of their work, they have a state waiver from some of the Regents Examinations). Research from their work indicates that New York City students who graduate from these schools (which have a much higher graduation rate than the City although they serve more low-income students, students of color, and recent immigrants) are more successful in college than students with a traditional Regents diploma which relies upon standardized tests.
- In Silicon Valley, CA, many school districts use the Mathematics Assessment Resource System (MARS), an internationally developed program which requires students to learn complex knowledge and skills to do well on a set of performance-based tasks. The evidence is that students do as well on traditional

tests as peers who are not in the MARS program, while MARS students do far better at solving complex problems.

- Australia, New Zealand, Hong Kong, Singapore, England, and Canada operate systems of assessment that include local performance-based assessments that count toward the total examination score (typically at least 50%). In Queensland, Australia the state's "New Basics" and "Rich Tasks" approach to standards and assessment, which began as a pilot in 2003, offers extended, multi-disciplinary tasks that are developed centrally and used locally when teachers determine the time is right and they can be integrated with locally-oriented curriculum. They are, says Queensland, "specific activities that students undertake that have real-world value and use, and through which students are able to display their grasp and use of important ideas and skills." Extensively researched, this system has had excellent success as a tool for school improvement. Studies found stronger student engagement in learning in schools using the Rich Tasks. Similar to MARS, on traditional tests, New Basics students scored about the same as students in the traditional program, but they performed notably better on assessments designed to gauge higher order thinking. The Singapore government has employed the developers of the Queensland system to focus their school improvement strategies upon performance assessments. High-scoring Hong Kong has also begun a process of expanding its already-ambitious school-based assessment system.

Clearly there is extensive experience available for designing and implementing assessment systems that include performance assessments, require multiple sources of evidence, and include local assessments. There is also an extensive research literature on performance assessments. The examples above are all examples of performance assessment *systems*; that is, assessment systems that use primarily or exclusively performance tasks, offering a strong existence proof for the viability of such systems.

Perhaps the most complex question surrounding these assessments when they are locally developed or scored is how to ensure comparability. Many of the systems described earlier, both in the U.S. and abroad, use common scoring guides. Queensland's system, like those in a number of countries, also employs "moderation," a process of bringing samples from different schools to be rescored, with results sent back to the originating schools. This process leads to stronger comparability across schools and is part of building a strong performance assessment system. The Learning Record, at one time used in dozens of U.S. schools, established very high inter-rater agreement (reliability) using moderation because the instrument is high quality and the training is effective.

Nebraska, through its peer review process, verifies that scorers within each district participate in extensive scorer training on common rubrics. Although districts may be using different tools, consistency and comparability within classrooms, buildings, and districts is supported in this way. Valid comparison across districts is achieved through external validation checks such as the statewide writing assessment, the ACT and other

commonly administered standardized tests. Each district's assessment system is evaluated and approved through a review process conducted by measurement experts.

Performance Assessment: Evidence

The research and work that has been done on performance assessment has uncovered a number of benefits, challenges, and criteria for making such assessment systems successful. Among the benefits of performance assessment systems are that they:

- Elevate the focus of instruction to higher order thinking skills;
- Provide a more accurate and comprehensive assessment of what students know and can do;
- Lead to more student engagement in both the learning and assessment process;
- Invite more teacher buy-in and encourage collaborative work;
- Support improvement of teaching practices;
- Provide clearer information to parents as to student development, accomplishments, and needs; and
- Allow instruction to be altered in a timely fashion to meet student learning needs.

From the research and evidence on performance assessment, there are a number of lessons learned that should be considered when designing a system that substantially incorporates performance-based assessments:

- Although some methods of managing performance assessments can cost more then machine scoring of multiple choice tests (i.e. when such assessments are treated as traditional external tests and shipped out to separately paid scorers), the cost calculus changes when assessment is understood as part of teachers' work and learning – built into teaching and professional development time. Much evidence suggests that developing and scoring these assessments is a high-yield investment in teacher learning and a good use of professional development resources. In addition, performance assessment systems are not necessarily more costly than accountability systems that rely upon standardized measures of achievement. For example, Nebraska, which utilizes a locally managed assessment system, spends only $.03 per child (or $9,000) on outside assessment contracts while Ohio, relying upon standardized measures, spends $50.00 per child (or $92,000,000). In most European and Asian systems, and in those used in several U.S. states, scoring of assessments is conducted by teachers and time is set aside for this aspect of teachers' work and learning. While teacher time to create and score the assessments can be substantial, these activities lead to more skilled and engaged teachers. In contrast, external standardized tests provide teachers with little guidance on how to improve student learning when they simply receive numerical scores on secret tests months after the students have left school. Hence the professional development that seeks to help teachers improve achievement in this system is under-informed and ineffective.

- Extensive professional development is necessary for educators to learn to build, use, and score assessments that will inform and guide their teaching. Few teachers now have that knowledge, but they can and will develop it when given the opportunity, as has been demonstrated in many systems. The system must engage the adult learners in curriculum alignment, performance task development, scoring processes, and data analysis so that they 'own' the system and do not feel bypassed. This includes developing a peer review, audit, or moderation system that provides for a feedback loop, checks on quality, and includes directions for staff development.

- Productive use of performance assessments, like proper use of standardized tests, should be aimed at revealing areas needing improvement and should lead to curriculum and professional learning supports rather than punishments. Only if schools or districts show themselves unwilling to take advantage of support should sanctions be undertaken.

- Personnel in departments of education and legislatures at the state and federal levels must understand that only classroom teachers can directly impact instruction and learning. Therefore, their task is to provide assistance to teachers to make the system work.

- Careful attention must be paid to the performance tasks. They should be developed in response to criteria that establishes the technical quality of assessments (including checking for bias and fairness), high proficiency standards, consistent administration of assessment, and opportunity to learn what is assessed. They should also be constructed to allow students with special needs and those who are learning English opportunities to demonstrate their knowledge appropriately.

<u>Performance Assessment: Federal Legislative Initiatives</u>

In the reauthorization of NCLB, consideration should be given to how federal legislation could support these more sophisticated forms of assessment that support students in developing higher order thinking and reasoning skills. Congress should provide support for states to design accountability systems that use multiple performance measures of student achievement that include locally administered performance assessments. To that end, we would suggest that legislative language capturing the following items be located in the reauthorization of NCLB.

1. Allow for and encourage the use of locally administered performance assessments as part of a balanced system for reporting on school and student achievement, in keeping with the existing requirement in Section 1111 (b) (3) (vi) that multiple measures be used to assess higher-order thinking and understanding.

2. Provide funding to states and localities to develop such systems that meet criteria which include:
    i. Assurance of the technical quality of assessments used for state reporting so that the evidence of learning derived from the classroom, school or district performance assessments is accurate, valid and reliable for the purposes for which it will be used;
    ii. Assurance that the assessments are valid measures of state standards as well as local curricula;
    iii. Assurance that assessment measures are free from bias;
    iv. Demonstration of validation and verification processes, such as peer review, assessor training, and moderation or auditing.
3. Appropriation of funds for any state that chooses to undertake the development of school based performance assessments, in an amount no less than $10 million per state and scaled to the size of the state, to support professional development activities for teachers and school leaders associated with developing, implementing, and scoring such assessments and integrating their results in plans for improving instruction. Such funds could also be used for states to work in collaboration in the design and validation of performance-based assessment systems, the development of performance tasks or other materials, and the design of professional development.

A fuller detailing of these proposals is available.

## Appendices:

1. Criteria for locally-based performance assessments to use in comprehensive state assessment systems.
2. Validation and Verification of Locally-based Performance Assessments
3. Performance Assessment: A Short Bibliography

# APPENDIX 1: Criteria for locally-based performance assessments to use in comprehensive state assessment systems

State proposals for funding in a grant or pilot project should ensure that the assessments they propose to develop and use meet the following criteria:

- are performance-based [see definition, below];
- assess higher order thinking skills [as required in current law – see definition below];
- provide multiple sources of evidence of student learning [see definition below];
- are locally-based [see definition below] – (This may include the use of tasks or assessments that are locally developed or locally-selected or adapted from a bank of tasks and used when appropriate for evaluating student learning);
- are fair and unbiased;
- are based on local curriculum as well as state standards;
- are able to be integrated with curriculum and instruction in schools and classrooms;
- provide timely, diagnostically-useful information;
- employ principles of universal design, while allowing adaptation to specific needs of students, particularly English language learners and students with disabilities;
- meet technical requirements of validity and reliability for the uses to which they are put;
- can be used to demonstrate progress toward proficiency; and
- are accompanied by or integrated with extensive professional development (and, professional development supported by the Act may be used to develop, use, and score locally-based performance assessments, provided the funds are not simply used for scoring large-scale assessments)

Performance-based assessment refers to assessments that evaluate applications of knowledge to real-world tasks. Such assessments may include, for example, students' oral or written responses to questions or prompts, as well as products such as essays or research papers, mathematical problems or models; science demonstrations or experiments; or exhibitions in the arts. They may be specific tasks they may be compilations of a number of such tasks within or across subject areas.

Higher order thinking and performance skills refer to the abilities to frame and solve problems; find, evaluate, analyze, and synthesize information; apply knowledge to new problems or situations; develop and test complex ideas; and communicate ideas or solutions proficiently in oral or written form.

Multiple sources of evidence (sometimes termed "multiple measures") involve different sources and kinds of evidence of student learning in a subject or across subject areas. Multiple measures allow multiple opportunities to demonstrate achievement, are accessible to students at varying levels of proficiency, and utilize different methods for demonstrating achievement.

Locally-based assessments may include both common assessments, which are assessments developed for use at the school or district level, and classroom-based evidence obtained from curriculum-embedded schoolwork by students.

# Appendix 2: Validation and Verification of Locally-based Performance Assessments

Local performance assessments, including classroom assessments, are commonly used in the instructional process in order to provide feedback to students and to improve instruction. When such assessments are used for accountability purposes they need to be *validated* as appropriately measuring the knowledge and skills they intend to measure and *verified* as being evaluated in non-biased, consistent ways.

There are several widely-used means that schools, districts, states, and other nations have developed to validate and verify the scoring of state and local performance assessments. These include:

- expert and peer review,
- concurrent validation studies and "benchmark checks,"
- assessor training and calibration
- external auditing, and
- moderation strategies.

We describe these methods briefly here and provide an example of how several of these strategies (peer review, benchmark checks, and assessor training) are used in the Nebraska assessment system, which relies on local assessment systems to complement the state's large-scale assessments.

## Validation and Verification Processes

Around the world, performance tasks, projects, and collections of student work – including the Advanced Placement and International Baccalaureate examinations – are used as part of both formative assessment systems and formal examination systems that carry accountability purposes. To ensure that the assessments themselves are valid measures of the intended learning standards and appropriately evaluate what students know and can do for the intended purposes of the assessment, they are typically subjected to several kinds of ***expert review*** – both of the tasks themselves and of the scoring tools and processes used to evaluate them. This review is typically conducted by experts in the content fields being assessed and by measurement experts and may draw on pilot studies and other research evidence about student performance on the assessments.
These reviews are a means of establishing content and construct validity for the assessments.

Another form of validation is to examine outcomes on assessments in relation to those on other measures. This is done through studies of concurrent validity, which are also sometimes known as "***benchmark checks***." If there are large discrepancies between the aggregate performances of students on different measures that are not explained by

differences in the skills and content they are measuring, this is a flag for further examination of how the assessments are being designed or scored.

Research shows that performance tasks can be scored with high levels of reliability if they are well-designed and guidance for scoring is clear and well-structured. This usually involves a rubric showing the scoring dimensions and descriptions of each performance level, along with instructions for how to evaluate the tasks. Consistency is greatly strengthened when the scoring guides are clear and of high quality. Collections of student work (work samples, portfolios) can be reliably scored when students and teachers have clear guidance on the features of the work to be submitted that facilitate consistent scoring.

***Training assessors*** also helps ensure that tasks are scored consistently and in an unbiased fashion. Assessor training typically involves learning the scoring process from an expert and reviewing benchmarks, which are assessment samples that represent responses at each score level (e.g., basic, proficient, advanced). Discussion of these examples helps bring scorers onto the same "page" – sharing a common agreement on what exemplifies work at a given level. The generally agreed-upon form of determining score consistency is inter-rater agreement: the extent to which raters agree with each other's scores. Agreement rates of .8 and above are seen as strong and generally adequate for most purposes. The supports for ensuring high levels of agreement are the proper selection of the materials submitted for moderation, high-quality scoring guides, and thorough training of the assessors. In some systems, those who are unable to regularly reach appropriate levels of agreement are not certified as assessors.

***Moderation*** is often used to establish reliable scoring, either as part of a training process or as part of the double scoring of tasks. Moderation is a process through which tasks are scored by two or more trained readers to help readers calibrate their judgments. Sometimes, moderation is used for tasks that are just at the "cut score" for a passing or failing grade. In such moderation sessions, especially if significant stakes are attached, two readers are assigned; and if they do not agree, a third, supervising reader makes the final determination. During scoring sessions, moderators may "drift" – for example, reading a series of especially good pieces may make a reader react too negatively to an average piece. To address these kinds of problems, in the stack of pieces a reader goes through there will be samples that have already been expertly scored (not revealed to the reader) so supervisors can check on drift.

Moderation results can be used to assign a final score or to provide feedback to teachers as part of a longer-term improvement process. This process has been used for both purposes in systems in the United Kingdom and in states such as Vermont. The Advanced Placement Art assessment also uses moderation to assign scores: trained judges score student artwork, giving each student his or her final AP score. And in many other AP courses, panels of teachers grade student essays. International Baccalaureate assessments, which are open-ended essays, projects, and products, are scored in a similar fashion.

The Learning Record, a system of assessment based on a tool developed in the U.K. to collect and evaluate samples of student work, uses moderation for long-term improvement. Only a random sample of Records from each participating classroom is re-scored. The scores given by the raters and their comments are returned to the originating teacher. While this does not change the score of any student, the evidence shows that, with feedback, teachers learn to evaluate their students more accurately.

***Auditing*** is a similar means of checking on the reliability of locally-scored assessments. This approach has been used for many years for the New York State Regents examinations which, like examinations in most European and Asian countries, are routinely scored by teachers in their local schools. A proportionate sample of tests is pulled and re-scored each year, and when a school's scores are flagged, they can be re-evaluated. If a school's assessments are not properly calibrated, additional training and guidance can be used to bring them in line. In some systems in other countries, such as the school-based assessment system in Victoria, Australia, school inspectors examine the tasks and student work samples that are scored locally and provide an overview of the quality of the work that is part of the feedback to the school and to the state agency for guiding the process of continual improvement.

## Validation and Verification of Locally-Based Performance Assessments: The Case of Nebraska

The verification and validation of locally-developed performance assessments in Nebraska is conducted with two primary considerations: a peer review balanced with technical expertise, and external benchmark validation.

### Peer Review and External Technical Expertise

In Nebraska each school district is visited on site by a knowledgeable and trained team of peers who are teachers or administrators in other school districts and who have experience in developing local performance assessment. The trained peers gather information about each local assessment system based upon a pre-determined set of technical assessment criteria. The review team examines the evidence available in the district and conducts conversations with local staff to determine the methodologies and processes used for establishing valid, reliably scored assessment, reviewed for fairness and appropriate level. In addition to examining the processes and the assessments, the school district must provide the validity documentation and reliability calculations, assuring that their processes have produced fair, accurate assessment results of sufficient quality for state reporting.

The Six Quality Criteria, developed in collaboration with the Buros Center for Testing at the University of Nebraska are as follows:

- The assessment items/tasks match the standards and are sufficient enough to measure the standards.
- The students are assured the opportunity to learn.

- The assessment has been reviewed for bias and insensitive language or situations.
- The assessment is at the appropriate cognitive level.
- The assessment is reliably scored.
- The mastery levels are appropriately set.

The peer review team gathers evidence from each district, but does not assign the final rating. That is left to a team of assessment experts, who are psychometricians. Each peer review team is assigned to an assessment expert. The expert and the peer review team discuss the information gathered, and draft collaboratively written feedback entered in an electronic data system for districts to receive in a timely manner. The final rating and any suggestions for improved processes are provided to the district by the technical external expert but in the language of practitioners. The validation processes provide opportunities for districts to visit with their peers, feel comfortable in sharing the evidence of their processes, and yet have the opportunity to receive understandable feedback (filtered through the peer review team) from the measurement experts.

Additional work that is required is noted, and a formal appeals process is implemented where districts indicate their intent to resubmit additional or clarified evidence within a department determined time frame. The department conducts a second review contracting with a balance of peers and the external assessment experts.

Training for the peer reviewers is extensive. The first round of training consists of two days prior to the review week. A second round of training occurs on the first full day of the review week. The training itself is a collaborative process facilitated by one expert Nebraska peer, the department of education, and one external psychometrician. In this way, the review teams have the opportunity to see the collaboration and balance of local review and technical expertise.

Scoring rubrics are detailed, thorough, and distributed well in advance to districts. These scoring guides include clear expectations by the Department of Education for the evidence to be provided. The scoring process includes orientation, practice scoring with the scoring rubric, and team scoring. Reviewers practice the scoring process with samples of district evidence of varying quality that have been selected for the training. A set of visitation guidelines are reviewed with all peer reviewers so that each district can experience a similar procedure.

***External Validation – Benchmark Checks***

Locally-developed assessments are not the only data source used to determine how well students are performing inside Nebraska school districts. Multiple data sources are used to not only report student performance but to serve as a source of validation, or an "audit" of local assessment processes. Among the external validation benchmark "checks" in Nebraska are the following:

- A statewide writing assessment - generated, administered, and scored on the state level to all students in grades 4,8, and 11

- A required national achievement test required once in the elementary, once in the middle school, and once in the high school
- ACT results
- NAEP results

Additionally, each year the department conducts validity studies tracking the large-scale reading, mathematics, and writing results over time. These external tests are then correlated with the local assessment results. In this state, locally developed classroom-based performance assessments are an important part of a balanced assessment system.

# Appendix 3: Performance Assessment: A Short Bibliography

*Information on state assessment systems:*

Nebraska

Gallagher, Chris. Reclaiming Assessment (Portsmouth, N.H.: Heinemann, 2007).

Nebraska Assessment web site at www.nde.state.ne.us/stars/

New York

NY Performance Assessment Consortium at www.performanceassessment.org

Wyoming

"Wyoming Steers Clear of Exit Exams," FairTest Examiner, January 2007. (www.fairtest.org/examarts/2007%20January/Wyoming.html) and http://www.k12.wy.us/Saa/WyCAS/archive/PubsPresent/Pubs/AssessmentHandbook.pdf

Multiple States

Darling-Hammond, Rustique-Forrester, & Pecheone, Multiple Measures Approaches to High School Graduation (Stanford University: School Redesign Network, 2005)

*Information on International Approaches:*

Queensland, Australia

http://education.qld.gov.au/corporate/newbasics/html/richtasks/richtasks.html

*Information on Performance Assessment Systems:*

Mathematics Resource Assessment System

http://www.nottingham.ac.uk/education/MARS/

Learning Record

http://www.cwrl.utexas.edu/~syverson/olr/olr.html, and

http://www.fairtest.org/Learning_Record_Home.html