A Brief on Performance-Based Assessment Technical Considerations From an International Perspective

December 2009

Nick Pinchok
and
Arie van der Ploeg

Great Lakes East Comprehensive Center at Learning Point Associates

1120 East Diehl Road, Suite 200
Naperville, IL 60563-1486
800-356-2735 • 630-649-6500
www.learningpt.org/greatlakeseast/

This work was originally produced in whole or in part by the Great Lakes East Comprehensive Center with funds from the U.S. Department of Education under cooperative agreement number S283B050012. The content does not necessarily reflect the position or policy of the Department of Education, nor does mention or visual representation of trade names, commercial products, or organizations imply endorsement by the federal government.

Great Lakes East is one of the 16 regional comprehensive centers funded by the U.S. Department of Education, and its work is administered by Learning Point Associates.

3754R_1/10

Contents

Introduction

Ohio's vision and goals for new forms of assessment, outlined in the Department of Education's 2007 *Toward a New Generation of Assessment and Accountability*, provide an exciting leadership opportunity for the state. Seeking to create a "world-class" educational system, and using performance-based assessment as one of the ways to fulfill that vision, Ohio hopes to incorporate a more "multiple measures" approach to its statewide assessment program. This approach will require new demands on teachers, though, and will involve administering, scoring, and analyzing high-stakes assessments in different ways. Ohio's leadership and vision in this area of education and assessment reform likely will be watched and scrutinized in the near future to determine if this is a better way to measure student learning.

Because of some historical criticism of the credibility of performance-based assessments, the Ohio Department of Education has requested that the Great Lakes East Comprehensive Center provide a brief on the current state of performance-based assessment with respect to how other countries have used this form of testing to the advantage of their teachers and students. Although implementing a large-scale, performance-based assessment system at the state level is not as easy and poses arguably more challenges than scanning and scoring thousands of multiple-choice bubble sheets, very credible approaches on how to bring this to scale can prepare Ohio for a state-of-the-art system. By acknowledging and planning for these assessment, psychometric, and professional development requirements, the Ohio Department of Education and its schools can begin to incorporate a more multiple measures approach to their instructional and accountability decision making, and hopefully better prepare Ohio's high school graduates for college and our rapidly changing world.

Why Performance-Based Assessment in the 21st Century?

I find great emphasis on problem solving, applications of principles, analytical skills, and creativity. Such higher mental processes are emphasized because this type of learning enables the individual to relate his or her learning to the many problems he or she encounters in day-to-day living. These abilities are stressed because they are retained and utilized long after the individual has forgotten the detailed specifics of the subject matter taught in the schools. These abilities are regarded as one set of essential characteristics needed to continue learning and to cope with a rapidly changing world.—Benjamin S. Bloom (1978, p. 578)

A wealth of research has emerged over the past several decades in the area of assessment. As with most fields of research, the breadth of analysis and findings is compelling, varied, conflicting, and passionate. Much of the research basis is heavy on statistical analysis, quantifying the results of measurements that serve many purposes. Many of our country's most reputable assessments historically have been given en masse (the Stanford-Binet Intelligence Test, SAT, and Iowa Test of Basic Skills, to name a few); therefore, the public and most educators have accepted that mass-scored, mostly forced-choice forms of assessment make sense, are reliable, and are valid for making decisions about students because they have been part of the education system we grew up with.

During the second half of the 20th century, a branch of assessment research expanded around more authentic, less standardized forms of testing. Research on formative assessment, as it is typically called today, centered around eliciting quality feedback from learners (Bloom, 1968; Bloom, 1971), teacher judgment (Sadler, 1989), and cognitive research (Pellegrino, Chudowsky, & Glaser, 2001). Bloom's Taxonomy of Learning demonstrated that there are varying levels of cognition, and representing those forms of cognition while progressing up each level of the Taxonomy scale is important when designing assessments for all students.

As the research expanded, and as more performance-based assessments were folded into accountability systems, it came into conflict with some research and psychometric aspects of reliability and validity. Many experts (Messick, 1989; Sadler, 1989; Shepard, 1993) helped define the technical terms and aspects of what goes into the dependability and credibility of formative and performance-based assessments (see the Technical Considerations section [pp. 12–16] for terms).

For instance, Sadler (1989) states:

> Reliability is usually (and correctly) said to be a necessary but not sufficient condition for validity, because measurements or judgments may be reliable in the sense of being consistent over time or over judges and still be off-target (or invalid). … Attention to the validity of judgments about individual pieces of work should take precedence over attention to reliability of grading in any context where the emphasis is on diagnosis and improvement. Reliability will follow as a corollary. (p. 122)

Models of Top-Performing International Performance-Based Assessment

Many countries' assessment approaches have proven effective from an achievement and instructional perspective, as has been exhibited by high rankings on international assessments like PISA, TIMSS and PIRLS. Other countries' assessment systems have been built to facilitate learning and to better understand how students can apply their knowledge. Both small- and large-scale assessments focus on student inquiry, encouraging the exploration and synthesis of research, asking questions, formulating new ideas, and defending positions. The following examples could serve as models to Ohio when considering how to build a performance-based system.

New Zealand

In recent years, New Zealand has performed very well on tests such as PISA (4th in Combined Science Literacy and 7th in Mathematics Literacy in 2006), impressive rankings for a nation as diverse, rural, and relatively poor as New Zealand. Jeffrey Smith, codirector of the National Education Monitoring Project (NEMP), which oversees the New Zealand's national assessment program, says that reformers "can make an impact here," mostly due to the small size of New Zealand and the respect given to his organization (J. Smith, personal communication, June 2, 2009). Inheriting an education and assessment system that had historically used a multiple-choice approach, the original codirectors of NEMP, Terry Crooks and Lester Flockton, traveled the world looking for strong examples of research-based assessment approaches to bring to their country. Formative assessment research and practices were beginning to emerge around the world and NEMP was designed to take advantage of some of those concepts.

According to Smith (personal communication, June 2, 2009), the country built their national assessment system based on the following three principles:

- Light sampling (approximately 1 percent of student population)
- Low stakes with no child reporting or school reporting
- Assessing what students know and can do through complex tasks and problems

Crooks and Flockton were most interested in finding out how students could apply what they learned and "do" with content they had studied, and performance-based assessments supported that approach. During academic Years 4 and 8 (similar to U.S. 8- and-12-year-olds) students are randomly selected to voluntarily take the national assessments. A combination of one-on-one-administered and small-group activities are performed with teachers scoring these assessments. Selected and trained teachers are released for six weeks to complete these activities. Some tasks are videotaped, some are done with paper and pencil, and some are completed online. Tasks are designed to be complex and are coupled with student interviews and surveys to gain insights into the students' thinking. Only half the tasks are reported on publicly, with those non-reported tasks administered again four years later to show growth.

Marking schedules and criteria for rubrics are built into the task design and development process through piloting processes. This allows future trained scorers the opportunity to look for proper sequencing and potential flaws of student performance prior to administration. Local teachers, NEMP staff, and national curriculum advisory panels all participate in this process. A new set of marking sheets and a scannable scoring system have been designed to collect and analyze over one million items of marked data for statistical analysis. Cross-marking with teams of 20 in facilitated discussions is employed to increase consistency and further enhance the validity and reliability of all scores. In addition, NEMP researchers have performed studies into the extent with which their national assessment tasks and scoring procedures measure types of thinking, such as critical, creative, reflective, and logical. Findings to date are encouraging, and additional research continues into the effectiveness of this approach.

Hong Kong

Committed to systemwide assessment reform since 2000, the province of Hong Kong has rigorously studied, learned from, and incorporated the research on formative and performance-based assessments into their large-scale assessment systems and processes. According to a 2004 presentation by Hong Kong's Education Commission ("Progress of Education Reform"), areas of assessment reform included increasing "the validity of public exams [through] more open-ended questions [and] more emphasis on application and problem-solving formats." Along the lines of this belief of increased validity, the country's policies, research, tools, and professional development all center around the key tenets of formative assessment: student-centered emphasis, sharing performance standards with all key stakeholders, the reduction of public examinations in favor of more alternative types of assessments (Education Commission, Hong Kong Special Administrative Region of The People's Republic of China, 2006).

The entire primary-to-secondary system in Hong Kong is being retooled in many ways. All grades are incorporating more Assessments for Learning as part of their accountability and reform initiatives. In conjunction with this move toward a multiple-measures approach for secondary students, the Hong Kong Certificate of Education Examinations is being phased out in favor of the Hong Kong Diploma of Secondary Education. The intent is to bring in more authentic tasks to school-based assessment (SBA) measurements to provide a more comprehensive view of student performance. Assessment development emphasizes fieldwork, portfolios, projects, oral presentations, and other measures Hong Kong has even taken the process a step further than most, training its classroom teachers on Rasch-model software applications to more thoroughly analyze their students' performance local, formative assessments that can aid in differentiating instruction more accurately with their more performance-based SBAs.

The Hong Kong Examination and Assessment Authority (HKEAA) has worked diligently to include many key stakeholders in the process of standard descriptor and assessment development (both items and processes). This long-term project has been underway, and full-blown implementation is scheduled for 2012. Emphasis on quality assurance has been high, with efforts made to train teachers on scoring reliably, the use of technology in scoring to increase efficiency, moderation processes, and double scoring and statistical analysis. "The HKEAA will devise the assessment criteria, exemplars and guidelines to ensure consistency of marking standard among

teachers. It will also organize training courses to enhance teachers' understanding of the SBA.… The HKEAA will also take follow-up action to help schools overcome any difficulties encountered in implementing SBA" (HKEAA, 2009).

Initial stages of implementation appear to be working. The HKEAA is working with other countries to learn how they assess in this manner and ensure comparability to optimize best-practices implementation and to benchmark their scores to others around the world.

Finland

Considered having one of the world's premiere education systems, Finland has used formative and performance-based assessments as part of its comprehensive approach to teaching, learning, and assessment for years. Much of the credit for strong performance on international tests has been given to the quality of teachers the country produces and maintains. The brightest of Finnish college students typically pursue teaching careers. Only 10 percent to15 percent of applicants are accepted for teacher education programs (Jakku-Sihvonen & Niemi, 2006), and future teachers are expected to receive a master's degree and participate in thesis projects before entering their first year in the classroom. The expectation is excellent teachers who have a real-world knowledge of the latest education research as it pertains to teaching, especially of students with special needs. There is no centralized "inspection" system as seen in other countries—it was disbanded in the 1990s—and the country relies more on analysis of local assessments, as well as international assessments and reviews, to determine its educational strengths, needs, and education policy development.

The country has no real high-stakes testing until its voluntary matriculation exams, which most students participate in. Sahlberg (2006) states:

> An important factor affecting the nature of teaching and learning in general upper secondary school is the nature of student assessments and school evaluation. Teachers assess the achievement of each student at the end of each course which means approximately five or six times per subject per school year. The National Matriculation Examination that students take after successfully completing all required courses is a high-stake examination and has therefore a visible affect on curriculum and instruction. Nevertheless, general secondary school can be characterized by having a strong focus on learning, creativity and various methods of studying rather than concentrating on passing tests and exams. (p. 14)

In addition, Sahlberg writes:

> Teacher professionalism and society trust in schools and teachers have protected the Finnish secondary education system from many consequential accountability policies that are common in the United States, England and Canada. Instead, national curriculum and evaluation strategies are designed according to intelligent accountability principles (Secondary Heads Association, 2003; Crooks, 2003; Fullan, 2005). Intelligent accountability in the Finnish secondary education context preserves and enhances trust among teachers, students, school leaders and education authorities in the accountability processes and involves them in the process, offering them a strong sense of professional

> responsibility and initiative. For example, vocational education performance-based assessments are based on collective judgment and feedback from teachers, employers and employees in tandem with the voice of the student. Intelligent accountability designs in Finland also require that evaluation and assessment leads to deep, worthwhile responses rather than bold statistics and technical reports. In many cases schools and teachers have access to the assessment evidence concerning their own school in order to track down the areas of improvement.… National sample-based assessments in lower secondary school together with continuous teacher-made classroom assessments provide well-founded and immediate feedback that promotes insight into performance and supports planning and decision making about what works and what should be improved. Indeed, the national Matriculation examination at the end of general upper secondary school is the only high-stake accountability measure in Finland. (p. 22)

Finland's is a data-driven system. Teachers value the formative assessment approaches and judgments they use, and schools work with the Ministry of Education, the Finnish Educational Evaluation Council, and other partners to ensure assessments and scores are valid, reliable, and comparable. The eighth- and ninth-grade "final assessments," for example, are yearlong processes of collecting multimodal evidence of student learning through a variety of tasks, projects, and tests. Students receive a numerical and descriptive score at the end. At the national level, samples of the ninth-grade assessments are taken and evaluated among a variety of other data (i.e., surveys) to determine success of students and equity. Rather than a large-scale assessment system with high accountability, there is more of a process of external evaluation of the school system that is done through sampling student and teacher work for quality and assessments being only a portion of the analysis. A testing system based on sampling that is managed and controlled by teachers parallels similar approaches in New Zealand and Hong Kong.

United Kingdom

The United Kingdom (U.K.) has had a national system of assessment for decades. Most of its higher stakes, upper-grade assessments are managed by the Qualifications and Curriculum Authority (QCA), which provides services and supports for assessment development and teacher-scored assessments. The General Certificate of Secondary Education (GCSE), a voluntary test to earn this designation, allows students to take courses and tests for university-entrance purposes. .

The QCA has had procedures in place for years to "make marking reliable" (Baird, Greatorex, & Bell, 2004). The use of exemplar scripts, scaling of examiners' markings, and coordination meetings to monitor examiners have been outlined in Code of Practice documents so that awarding bodies distribute certificates with authority. Having networks for assessors, similar to our professional learning communities, to tap into is important for ongoing conversations and monitoring of quality work as well as consistent scoring. Examiner knowledge of standards, and how student work represents those standards, factors into the quality of the inter-rater reliability, and significant training and planning need to go into the system upfront to increase reliability and comparability. Sampling student work and "double scoring" has been a consistent practice for years.

Cambridge International Examination System, based in the U.K., has been in effect since 1853 and is used in more than 150 countries. Considered the standard for large-scale assessments, this methodology (with the organization based in England) has been adopted by countries all over the globe to measure and recognize quality performance by their students. New Zealand and Singapore, for example, offer Cambridge assessments as options for their high school students. With choices of syllabi and assessments in various subjects, these "external" measures of attainment have the connotation of a global benchmark of rigor *for* students, and attainment *by* students. Certificates earned on these assessments signify proper qualifications for colleges and universities internationally. These course certificates, typically identified as Levels A to O (A being the highest level), have held up over the years, although they continue to grow and change as needed.

More performance-based assessments have been added to their large-scale assessments over the decades. One type of addition includes the recent Applied Subjects approach to emphasize and measure application of knowledge for those students likely going into some kind of workforce-readiness program of study after high school. Other simulated or actual performance tasks are assessed by qualified examiners and are administered, scored, and analyzed through highly standardized processes. The development and scoring of tasks pass through formalized design (aligning purpose, items/tasks, and inferences), validity (minimize construct under-representation and construct-irrelevant variance), reliability (understand and document limitations of precision) and administration (from optimizing conditions through a fair appeals process) principles.

Australia

Many of Australia's states have implemented a performance-based approach to assessments for years and have been held up as examples of quality systems. Overall, this very diverse nation has performed respectably well on international assessments, especially in comparison with similar and less diverse, more affluent nations. Much of the success has been credited to an innovative approach to performance-based assessment.

As Cumming and Maxwell (2004) reported:

> Assessment reform at the upper secondary level has developed in Australia over the last 30 years for at least two reasons: first, to reduce the curriculum control exercised by universities through externally set examinations, in order to address the needs of the majority of students; and, second, to respond to the need to broaden the curriculum to include more practical and contextualized learning. Two additional themes emerging from this reform are: first, the preference for standards-based (or criteria-referenced) assessment reinforcing the work undertaken in developing an appropriate and broad school curriculum; and second, respect for teacher professionalism in judging student achievement. (p. 93)

Nearly all states in this country have adopted a similar philosophy, but with the power to decide on the specifics. Victoria, for one, is considered a model state for managing assessments at the teacher, school, and state levels. No state, however, has the history, documentation, transparency, and management that can compare with the Queensland state approach. For decades, Queensland

schools—in partnership with the Queensland Studies Authority (QSA) and other important entities such as the Australian Council on Education Research (ACER) and local universities—have adhered to a world-class system of writing, aligning, implementing, training, scoring, validating (moderating), and analyzing their performance-based assessments.

In Queensland, schools and districts start with course syllabi, which are approved by the QSA to ensure acceptable levels of rigor and consistency across the state. These syllabi include a broad range of skills with complex cognitive aspects, as well as assessment tasks. Assessment criteria are standards based, not norm referenced, so any percentage of students can obtain mastery or proficiency. Based on a wide range of formative assessment research, locally developed assessments are graded by teachers, and students are apprised of their performance along the way in a real-time, student-centric, evidence-based system. A system of Chief Moderators along with a Board of Senior Secondary Studies work in tandem with other partners to verify and authenticate the work (course or subject syllabi) that will be done by students. The externally "moderated" review process gives legitimacy to the scores and credibility to all key stakeholders**.** About 10 percent of teachers statewide participate in moderation. This system works well and consistently because there has been significant, systemic training done with teachers and principals to become familiar with the approach and to increase the knowledge and skill surrounding what formative assessment is and how it works.

In addition, the Queensland Core Skills (QCS) test acts as an equating measure to provide evidence of alignment to the school-based assessments. Ironically, the school-based assessments have carried the most weight in making final judgments of students.

As a model of transparency and technical soundness, the QSA and its partners look closely at the factors that impact psychometric quality, and they perform ongoing studies and analysis of their moderation techniques, sampling of assessment items, and comparability. The reports are posted online and open to the public. Many factors influence the successful implementation of a large-scale performance-based assessment system, such as the development of clear, articulated learning targets, and quality professional development with an eye on good instructional practices related to assessment results. The defensible results associated with Queensland's approach deserve continued attention and study.

Summary of Technical Aspects of the Large-Scale Systems Overseas

Great Lakes East attempted to analyze the psychometric principles shared by some top-performing countries, states, and provinces overseas with respect to how they administer, score, analyze, and defend their large-scale, mostly performance-based assessments. The data collection consisted of a matrix template designed in conjunction with multiple well-known assessment experts. This template asked respondents for descriptions, or examples, of the following technical aspects involved in their large-scale assessment programs:

- Goals and purpose
- Data collection
- Scoring and analysis

- Psychometric principles
- Reporting of scores
- Embedded professional development

Great Lakes East requested voluntary, unpaid submissions from key staff members either charged with overseeing testing and assessment departments in each of the countries, states, or provinces or from external staff members related to the running, researching, or supporting of those systems (i.e., university members who have done research or development for specific QSA's or Ministries of Education). Eleven individuals were contacted to respond representing the following:

1. Cambridge Assessment
2. England
3. Finland
4. France
5. Germany
6. Hong Kong
7. New Zealand
8. Norway
9. Queensland (AU)
10. Scotland
11. Victoria (AU)

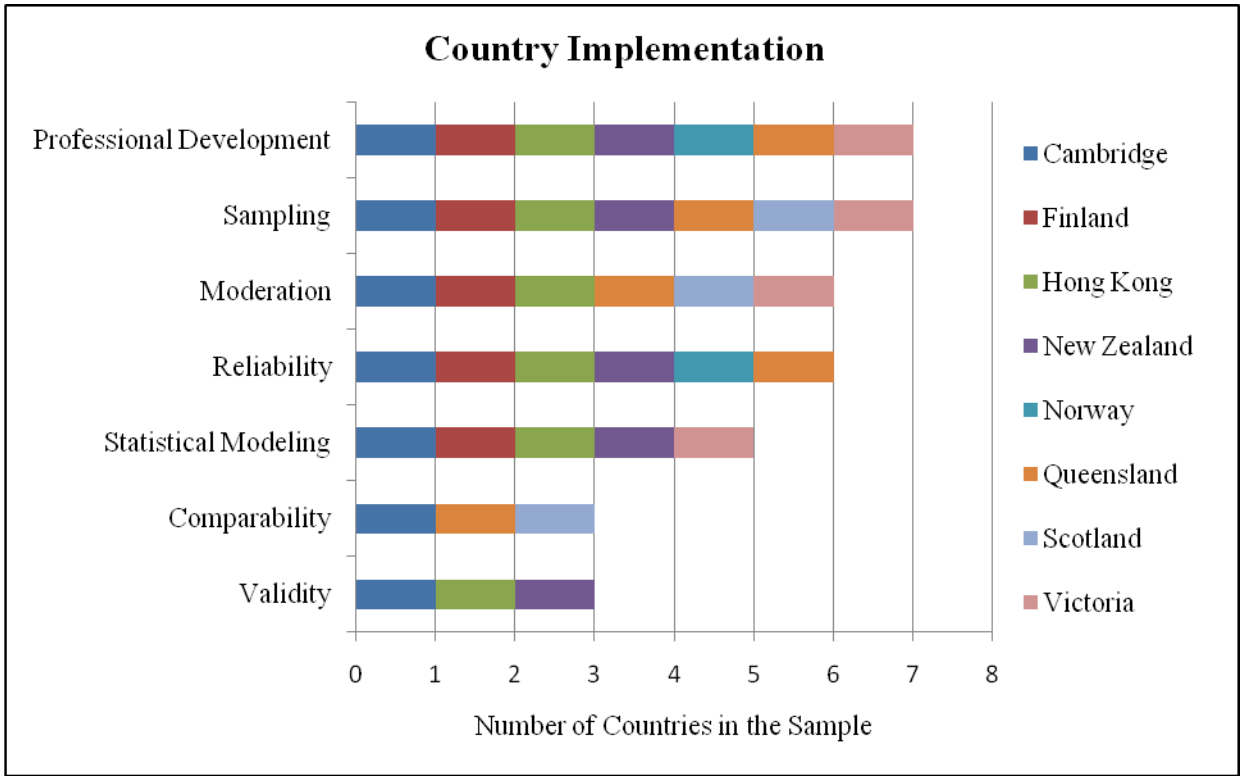
In total, eight of the 11 potential contributors responded with completed templates or shared supporting research or papers they had written on their locale's testing system. Their submissions were analyzed for common themes and technical aspects that would contribute to a defensible psychometric framework for large-scale, mostly performance-based assessment systems.

Matrix content and supporting documentation were explored primarily for the prevalence of, and reliance on, the following psychometric and quality control aspects of the large-scale systems:

- Comparability
- Construct validity
- Moderation processes
- Professional development in the administration/scoring of large-scale assessments
- Reliability
- Random sampling effects
- Sophisticated statistical modeling and analysis

The submissions were not intended to be definitive or exhaustive summaries of each entity's systems, and the analysis did not include information obtained from official technical manuals (Great Lakes East was unable to attain any). Figure 1 portrays the coverage of the technical aspects (listed above) across eight countries.

Figure 1. Prevalence of Technical Aspects of the Large-Scale Systems in Eight Countries



Based on this sampling analysis, *large-scale assessment systems overseas place a higher emphasis on reliability, teacher scoring and moderation, sampling, and professional development to achieve their technical rigor and defensibility*. They also value validity, sophisticated statistical modeling, and comparability, but not as widely as the previous traits/aspects.[1] (See the Appendix for a more thorough analysis of participating countries' responses.)

It is clear that nearly all the countries, states, and provinces surveyed and those who responded have not worked within a system that requires assessment to be based mostly on accountability as a driving factor. The following comments from participants' submissions exhibit that other nations have attempted to use large-scale, performance-based assessment as a tool for improving learning more so than determining student "proficiency" or for teacher and school accountability:

- Scotland: "Recognize the importance of the professional judgment of teachers with respect to classwork scoring validity."

[1] Though not an official, complete study, it is both a summary and tool to identify patterns in approaches across these countries. Great Lakes East hopes to participate in a more thorough analysis, with well-respected, cooperating research partners in the very near future.

- Queensland: "Reports of student achievement are defensible and comparable, based on sound evidence and shared understanding of Essential Learnings and the scope of syllabuses."
- Hong Kong: "Each teacher in Hong Kong is entitled to 150 hours of free professional development in three years."
- New Zealand: "NEMP is very clear that it cannot measure what schools have taught children: The capabilities, attitudes, and preferred activities of children reflect all of their life experiences, not just what they have learned in the limited portion of their lives spent in school."

Crooks reinforced the justification of these different driving technical qualities. Through a series of both phone and e-mail correspondence with Great Lakes East, Crooks explained that "where the assessment is done in ways that sideline the teachers, the potential for teacher learning and professional development is greatly reduced, and the incentive for teachers to try to 'game' the system is increased." In addition, he acknowledged:

> …the tension between reliability and validity…. Most psychometric efforts focus on maximizing reliability, because they have powerful statistical technology for doing that. Validity is much harder, because it involves human judgment, so [it] tends to get lip service rather than profound effort. Where validity is focused on, the range of assessment tasks and approaches tends to be broader, reducing the correlation between tasks and therefore the measures of reliability…. This is easier to get away with for national monitoring, where the reliability of individual student scores is not an important issue, but is very hard to get away with where the scores of individual students are central to the process. (T. Crooks, personal communication, 2009)

Though improving learning and determining effectiveness seem to be complementary, assessment as learning and assessment for learning can alter the methods of, and psychometrics around, tests, items, and gathering evidence. The thoughtful design concepts and investments in staff expertise appear to be critically important in a defensible and comparable performance-based assessment system.

Much can be learned from these international examples and approaches. Crooks finally comments:

> If we could all move towards Finland's model—less testing, less accountability pressure, and deep community respect for teaching as a profession and teachers as individual—we would go a long way to achieving our educational goals.... Whatever we try to do in assessment should always begin with the question: What is this going to do to motivate, inspire, and guide teachers and learners towards best practice, and what are the risks that it will do the opposite? How do we try to ensure the benefits while minimizing the damage? (T. Crooks, personal communication, 2009)

Hopefully, the defensibility requirements for Ohio, with the proof that decisions based on multiple measures are more reliable than singular testing events behind them, can hold up with Ohio's vision for a more innovative 21st century generation of assessments.

Defensibility of Performance-Based Assessments as Large-Scale Assessments

High reliability sells tests. When it comes time to select among competing standardized achievement tests, the decision makers (say, a district or state test-selection committee) will look to many evaluative factors to determine which test is best.... Test developers diligently seek high indicators of reliability and the score-spread that helps create such high reliability. Substantial score spread not only contributes to more accurate discriminations among examinees, it also helps peddle tests. And well-peddled tests make more money for the shareholders in the corporations that build and sell standardized tests.—W. James Popham (2001, pp. 46–47)

Interest in international performance and benchmarking has been building recently. It's fair to acknowledge that many of the highest performing countries place a priority on performance-based assessments. Even with the complementary research supporting the validity and fairness of more balanced approaches to large-scale assessment (Pellegrino et al., 2001), including performance-based assessments, there still seems to be some resistance to moving toward this approach. Though multiple-choice assessments can be highly valid and reliable, as well as less expensive and faster to score and report on, performance-based assessments do appear to be an enhanced way to measure the higher-order thinking skills that many education experts, and Ohio, are calling for. Can these performance-based assessments be brought to scale in Ohio, and nationally, in a defensible way?

The longest and arguably most comprehensive U.S. performance-based assessment system, Maryland's School Performance Assessment Program (MSPAP), had well-documented psychometric qualities (Yen & Ferrara, 1997). External analysis showed respectable-to-high levels of measurement accuracy, including but not limited to coefficient alphas (reliability), standard errors of measurement, and inter-rater reliability. Maryland's former director of assessment and the man responsible for building the system, Dr. Bill Shafer, believes that the advantage of this approach is that the "transparency of the test made it possible to see what good performance looks like," and it led to "good reliability of the test" (B. Shafer, personal communication, April 16, 2009).

Shavelson et al. (2003) studied multiple models of large-scale assessment systems that attempted to marry formative and summative assessment functions. The analysis looked closely at the designs of three systems: California's Learning Assessment System (CLAS), the U.K.'s proposed Task Group on Assessment and Testing (TGAT), and Queensland's (Australia) Senior Certificate Examination System. The first two models were not sustainable because of various complicating factors, but they had compelling attributes that would have contributed to quality systems. From the start (nearly four decades ago), Queensland's intent was to build a system that incorporated the best of formative assessment research into a summative examination. Born out of the Radford Committee report, Queensland replaced the British A-Level external examination with a school-based system which emphasized performance-based tasks and relied on teacher judgment for reliable scoring. Perceptions at the time were that both the public and teachers would be overly concerned with what was on the test, and that teachers would teach to the test and schools would narrow their curriculum (syllabi) to increase scores.

The Queensland approach had excellent curriculum alignment/design, implementation, verification (moderation), and psychometric (scaling, equating, comparability, etc.) aspects. Viewed favorably in the Shavelson et al. (2003) report from both a technical and instructional/pedagogical perspective, the proof of the analysis has been the long-term viability and sustainability of Queensland's approach. It is three decades old and continues to adapt and improve over time as more research and best-practices data come into the system and develop internationally. Shavelson et al. (2003) concluded that this performance-based approach is "so interactive [and] is a key component of accountability, supported by 'scientific evidence.' (e.g., Black & Wiliam, 1988)" (p. 36). And more traditional types of summative assessment must "not be allowed to go on limiting learning" (Shavelson et al., p. 36).

One challenge faced by performance-based assessments may come from the field's technical manuals. For the most part, standards for test technical manuals have been set by off-the-shelf, norm-referenced tests and test publishers. These publishers also have been involved with, or produced, the more recent states' high-stakes test documentation. A recent review (Pellegrino & Marion, 2006) of the literature was done with an eye on alignment to alternative assessment. Most alternative assessments for students with disabilities have performance-based aspects to them, yet the technical traits (i.e., construct validity) of these tests have not been well documented. This review aligned the research around formative assessment and the *Knowing What Students Know* (Pellegrino et al., 2001) three pillars (cognition, observation, and interpretation) with the technical needs of more performance-based forms of assessment for high-stakes purposes. Pellegrino and Marion (2006) argue that an evaluation perspective versus a strictly statistical/ psychometric one makes sense. Shepard (1993) pondered whether performance-based assessments improved the quality of work of all students, which most anecdotal and international evidence supports, and that the technical analysis should focus mostly on the "construct, relevance, interpretation and social consequences" of these assessments. From this viewpoint, Pellegrino and Marion (2006) state that more policies and technical analysis are needed to determine if all the evidence (theoretical, logical, and empirical) builds an argument for or against this form of assessment rather than if the items are simply valid.

As stated in *Knowing What Students Know* (Pellegrino et al., 2001), observation refers to "a set of specifications for assessment tasks that will elicit illuminating responses from students" (p. 2). That is very difficult to do simply from paper-and-pencil, bubblesheet responses. This is a potentially seismic shift in technical documentation and approaches if one values these arguments. Haertel (1999), referring to how many aspects of traditional, large-scale assessment documentation are stand-alone collections of analysis, stated that individual pieces of evidence do not collectively make an assessment system valid. Pellegrino and Marion (2006) state that "the model of how students develop proficiency in the domain … must be aligned with the methods used to collect observations and interpret those observations" (p. 53). The argument that schools over a century ago were designed to serve an agrarian-based economy but have not changed radically in design and, hence, do not presently serve how students learn best and how the brain works is similar to an argument about large-scale assessments: Most are designed for efficiency in scoring and in the psychometrics that support those scores rather than for how best to measure learning and elicit evidence of cognition and applications of learning. Pellegrino and Marion's analysis should be a new model for documenting performance-based assessment and for proving that, moving forward, these forms of assessments are observable and statistically valid and reliable.

Technical Considerations for the Ohio Performance Assessment Pilot Project

Validity

Though the definition of validity continues to be debated to this day, its importance to performance-based assessments is critical. The top priority in design and analysis is defining the construct so specifically that it can be operationalized and measured with consistency. As Messick (1989) states, "The validity of score interpretation and use depends on the fidelity between the constructs being measured and the obtained scores." Quality assessments need to gauge how well the underlying task's construct is measured and ensure that rubrics are defined and delineated, and that training is clear on the adherence to the descriptor criteria.

You will want to avoid "threats to" construct validity through construct underrepresentation (failing to capture critical aspects of the construct) and construct-irrelevant variance (elements that are irrelevant to the assessed construct) (Messick, 1994). Irrelevant score variance will occur if the task is not relevant to the construct and is often impacted by longer answers. Increasing the number of quality tasks increases the validity while decreasing the construct irrelevance and construct underrepresentation.

Crooks, Kane, and Cohen (1996) developed the following eight-stage model (see Table 1), based on Messick's construct validity framework, to view potential threats to validity:

Table 1. Eight-Stage Model

Stage	Threat
1. Administration of assessment	Low motivation, assessment anxiety, inappropriate assessment conditions, task or response not communicated.
2. Scoring of student's performance on tasks	Scoring fails to capture important qualities of task performance; lack of intrarater or interrater consistency; scoring too analytic; scoring too holistic; undue emphasis on some criteria.
3. Aggregation of scores on individual tasks to produce one or more combined scores	Aggregated tasks too diverse; inappropriate weights given to different aspects of performance.
4. Generalization from particular tasks included in combined score to the whole domain of similar task (the assessed domain)	Conditions of assessments too variable; inconsistency in scoring criteria for different tasks; too few tasks.
5. Extrapolation from the assessed domain to a target domain containing all tasks relevant to the proposed interpretation	Conditions of assessment too constrained; parts of the target domain not assessed or given little weight.
6. Evaluation of the student's performance, forming judgments	Poor grasp of assessment information and its limitations; inadequately supported construct interpretation; biased interpretation or explanation.
7. Decision on actions to be taken in light of the judgments	Inappropriate standards; poor pedagogical decisions.

Stage	Threat
8. Impact on the student and other participants arising from the assessment process, interpretations, and decisions	Positive consequences not achieved; serious negative impact occurs.

Comparability

In order to measure groups of students over time, it is important to carefully design tasks with an eye on articulated content and skills, scored with consistent procedures. Having equated forms (with comparable content and outcome scores) across the state is important and needs to be built into the design. It will be important for Ohio to embark on a public comparability study that is evidence based Traditional comparability studies entail the following: (1) an analysis of the "demands" placed on students by the syllabus, marking scheme, and question paper; (2) cross-moderation processes to compare work and determine which process has the better quality; or (3) a statistical analysis of award outcomes in syllabi/specifications under consideration (Yim, Shaw, & Lewis, n.d.). Historically in the U.K. (Elliott & Greatorex, 2002), comparability studies have used three different analytic methodologies: (1) Home and Away involves "home" examiners and "away" (external) examiners who judge student work. Results are analyzed using Kendall's coefficient of concordance to determine accurate scoring. (2) Thurstone Pairs is a mix of judges who analyze from a sampling which "script" or work is better. All the judgments are combined and analyzed using Rasch analysis to rank and determine different standards of student work. (3) With Kelly's Repertory Grid, senior/expert examiners define traits of quality work, evaluate their collective constructs, and distribute a questionnaire to other examiners who complete it prior to the comparability study. Statistical comparability analysis attempts to determine if similar candidates receive similar scores. Performance indicators/standards must be the same/similar to perform such analysis.

Moderation Techniques/Auditing

Quality moderation ensures comparability in order to standardize on performance-based assessments beyond interrater reliability; especially for accountability and comparability purposes, a system of moderation must be implemented. This approach can enforce and account for judgment and provide an empirical, standardized system for those judgments. It can start with peer-review teams and scale up with external modification/auditing. The designation of experts will be critical, and key partners in this process will need to be identified (universities, regional offices, etc.). The certification process of scorers and moderators/external auditors should be thorough and standardized. Successful countries, such as Finland and Australia, develop this expertise locally, with Finland incorporating most of the assessment skills, knowledge, and competency into its teacher preparation programs. A suggestion is to study the generalizability effects of committee error using generalizability coefficients of various facets.

Reliability

In performance-based assessment scoring, it is important to find the correlation between two observations of the same measure. The "rater-mediated" nature of performance assessments tends to lead to analyzing aspects of scoring for error rather than items. Inter-rater reliability

factors, such as training and calculating coefficient alphas, will be important. Beyond simple error, other factors can include sampling, equating, and whether the state wants to incorporate a system that allows multiple attempts for the student to be assessed and get to mastery.

The proportion of variability in the measure is also important. Engelhard (2002) states that construct-irrelevant variance factors include halo effect, differential interpretation of the score scale, gender-influenced ratings, and bias toward task difficulty. Teacher, scorer, and moderator familiarity with task, content, construct, and standard also lend themselves to variance. In Baird, Greatorex, and Bell's (2004) "What Makes Marking Reliable," several key factors were noted for increasing reliable scoring of performance-based assessments:

- Monitoring the order of marking (scoring).
- Establishing networks of examiners to discuss their work.
- Using standards of student work versus just criteria, especially with multiple criteria.
- Scoring work twice when possible.
- Training staff on the use of criteria and standards.

Scaling

When developing the range and difficulty of items for performance-based assessments, it is important to place additional weight on more reliable tasks/items. This calibrating can help to increase many factors that go into test defensibility. Scaling is typically done through text statements (i.e., performance descriptors), with numbers assigned to them on an interval scale. There may be a need to build in partial-credit to analysis. Maryland used a Masters' Partial Credit Model (two-parameter partial credit, same model used for NAEP) on MSPAP. Other forms of statistical analysis that can be done are for IRT, task difficulty estimation; linking the predictive, scale aligning and test equating aspects; and the interchangeability (same constructs, same levels of difficulty and reliability, different populations) of the items. It is critical that the scale score integrity be maintained over time.

Bias

It is desirable to elicit equivalent and/or similar responses and cognitive levels of activities from all students. With increased bias, there is reduced construct validity. Gender, racial/ethnic, and other forms of bias need to be eliminated.

Differential item functioning (DIF) analysis needs to be performed in order to eliminate bias from the testing process and identify problematic tasks, content, and processes/activities. Thought more difficult to do with performance-based assessment than with multiple-choice formats, a thorough item analysis of tasks and items should be performed regularly to reduce bias and calibrate difficulty and discrimination. Training of item writers is critical to reduce random and systematic error. Incorporating moderation techniques can also reduce or eliminate bias from teacher scoring and judgments.

Conclusion

The field of international benchmarking—specifically, determining how large-scale, performance-based assessments impact learning—is still in its infancy. Countries that implement this approach, however, have accumulated, in some cases, decades' worth of data. The Ohio Department of Education's choice to pursue a more multiple measures system of large-scale assessment and to use performance-based assessment appears to have support from these international benchmarking efforts. This type of assessment system, if designed correctly, can lead to broader and deeper measures of what students know and can do with their knowledge in defensible and comparable ways.

Clearly, more work needs to be done, as Lane and Stone (2006) request:

> Additional research on how students acquire and develop knowledge and skills within a content domain is needed so that performance assessment can better reflect domain-based theories of achievement and learning and better emulate the performance of interest. In turn, additional work is needed in identifying techniques for designing performance assessments and scoring procedures that capture these domain-based theories of achievement and learning. Developments in computer-based assessment systems that capture cognitive models of achievement and learning are beginning to emerge and hold promise for the future design of performance assessments. (pp. 423–424)

The research and recommendations outlined in this document are not based on a formal study; rather, they are based on an analysis of the current literature and resources available, including interviews with current and past assessment directors of large-scale performance-based assessment initiatives. A more thorough study of the global field would be welcomed, informative, and timely because of the current interest in international benchmarking and U.S.-based, high-stakes assessment reforms in progress. Great Lakes East at Learning Point Associates is interested in participating in such studies and hopes to be involved in any activities the Ohio Department of Education wishes to pursue along these lines.

References

Baird, J., Greatorex, J., & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education, 11*(3), 331–348.

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment (UCLA-SCIEP), 1*(2), 1–12.

Bloom, B. S. (1971). Mastery learning. In J. H. Block (Ed.), *Mastery learning: Theory and practice*. New York: Holt, Rinehart and Winston.

Bloom, B. S. (1978). New views of the learner: Implications for instruction and curriculum. *Educational Leadership, 35*(7), 563–576.

Crooks, T., Kane, M., & Cohen, A. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy, and Practice, 3*(3), 265–286.

Cumming, J. J., & Maxwell, G. S. (2004). Assessment in Australian schools: Current practice and trends. *Assessment in Education, 11*(1), 89–108. Retrieved June 16, 2009, from http://cmap.edu.fi/servlet/SBReadResourceServlet?rid=1G5NHJ2KK-1NV5F2G-2BS

Education Commission, Hong Kong Special Administrative Region of The People's Republic of China. (2006). *Progress report on the education reform. Learning for life; leaning through life.* Hong Kong: Author. Retrieved June 16, 2009, from http://www.e-c.edu.hk/eng/reform/Progress%20Report%20(Eng)%202006.pdf

Elliott, G., & Greatorex, J. A. (2002). A fair comparison? The evolution of methods of comparability in national assessment. *Educational Studies, 28*(3), 253–264.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–287). Mahwah, NJ: Erlbaum.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5–9.

Hong Kong Examinations and Assessment Authority. (2009). *HKCEE history: Guidelines on school-based assessment.* Retrieved June 16, 2009, fromhttp://www.hkeaa.edu.hk/DocLibrary/SBA/CE-Hist-09Guide-Eng-0809.pdf

Jakku-Sihvonen, R., & Niemi, H. (Eds.). (2006). *Research-based teacher education in Finland: Reflections by Finnish teacher educators.* Helsinki: Finnish Educational Research Association.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–431). Washington, DC: American Council on Education.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academies Press.

Pellegrino, J. W., & Marion, S. F. (2006). Large scale alternative assessment (A validity framework for evaluating the technical quality of alternate assessments). *Educational Measurement: Issues and Practice, 25*(4), 47–57.

Popham, W. J. (2001). *The truth about testing: An educator's call to action.* Alexandria, VA: Association for Supervision and Curriculum Development.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144.

Sahlberg, P. (2006). *Raising the bar: How Finland responds to the twin challenge of secondary education?* Washington, DC: World Bank. Retrieved June 16, 2009, from http://www.ugr.es/~recfpro/rev101ART4ing.pdf

Shavelson, R. J., Black, P. J., Wiliam, D., & Coffey, J. (2003). *On linking formative and summative functions in the design of large-scale assessment systems.* Manuscript submitted for publication. Retrieved June 16, 2009, from http://www.stanford.edu/dept/SUSE/SEAL/Reports_Papers/On%20Aligning%20Formative%20and%20Summative%20Functions_Submit.doc

Shepard, (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education* (Vol. 19, pp. 405–450). Washington, DC: American Educational Research Association.

Yen, W. M., & Ferrara, S. (1997). The Maryland School Performance Assessment Program: Performance assessment with psychometric quality suitable for high stakes usage. *Educational and Psychological Measurement, 57*(1), 60–84.

Yim, W. K., Shaw, S., & Lewis, M. (n.d.) *A science comparability study between two exam boards using a rank-ordering methodology at syllabus level.* Retrieved June 16, 2009, from http://www.aea-europe.net/userfiles/Bulgaria%20poster%20file%20Yim.pdf

Appendix
Responses of Participating Countries

	Cambridge	Finland	Hong Kong	New Zealand	Norway	Queensland	Scotland	Victoria
Comparability	Incorporate multiple comparability techniques	No data reported	No data reported	No data reported	No data reported	Externally moderated system ensures comparabil-ity by systems designed to reach the same standards across state	Compare teacher judgments across National Assessment bank score (when teacher determines students have reached a certain level of proficiency)	No data reported
Moderation Techniques	External moderation	10% of student papers are checked by external readers	Combination of school-based and external moderation	No data reported	No data reported	Externally moderated school-based assessment system since 1972; appeals process at the state review level if district review panel leads to	Moderation done at the school level by local authorities	Use statistical moderation to match levels and the spread of scores from school-based assessment and other external, more

	Cambridge	Finland	Hong Kong	New Zealand	Norway	Queensland	Scotland	Victoria
						disagreement		multiple-choice measures
Professional Development	Engage in significant administration, scoring, and quality-assurance professional development	Nationally recognized statisticians educate staff on the interpretation of scores	Teachers receive up to 150 hours of state-provided PD in areas of assessment of/as/for learning, Rasch modeling, and scoring to support system quality	Significant investments in training are involved in preparing staff for six-week process of administer-ing and scoring NEMP assessments	Significant effort is put into providing national professional development in the area of formative aspects of the national assessment	Trained moderation panels at the school, district, and state levels receive specific training to earn expert- or review-panel status	No data reported	Markers undergo full day of training to be certified
Reliability	Quantify/reduce measurement error while documenting steps to increase reliability while acknowledging (human) limitations	Mainly classical Alpha model in scoring and estimating reliability	Strong emphasis on reliability at the teacher and system level	Coefficient Alpha is tracked and cross-marking teams ensure reliability for consistency of scoring; reliability is sacrificed for higher levels of validity, though	Cronbach Alpha is tallied, and *p* values are calculated during development phase	Moderation processes ensure the highest checks on reliability of scoring and are monitored as such	No data reported	No data reported

	Cambridge	Finland	Hong Kong	New Zealand	Norway	Queensland	Scotland	Victoria
Sampling	Sampling used systemically to ensure expert judgment quality remains high	Use stratified and clustered samples nationally (10% of schools)	Sampling of students used in equivalents of Grades 3, 6, and 9 to monitor systemwide performance	Use 1% of students to determine national performance	No data reported	Queensland Studies Authority randomly samples approx. 10% of student portfolios annually	Assess only one content area per year with a percentage externally marked	Sampling new scoring methods to determine future direction of system
Statistical Modeling	Multilevel modeling used but acknowledge challenges of outliers	Scores are equated with earlier tests; final scores are equated as % of maximum score; use item discrimina-tion with Classical modeling	Rasch modeling additionally with EFA and CFA	Multivariate analysis	No data reported	No data reported	No data reported	Rasch analysis as student logit scores are trans-formed after linking between years to create scale scores
Validity	Cataloguing of constructs; minimize construct-irrelevant variance; sample adequately across subject and performance/ process domains	No data reported	Strong emphasis placed on validity in system	Validity is prime consider-ation, both construct and content; enhanced by making tasks motivating for students	No data reported	No data reported	No data reported	No data reported