Upgrading High-Stakes Assessment

Albert Oosterhof

March 2011

Center for Advancement of Learning and Assessment
Florida State University, Tallahassee, FL
www.cala.fsu.edu

Acknowledgments

The work reported in this paper is supported through a grant from Education Research Programs at the Institute of Education Sciences (IES), award number R305A110121, administered by the U.S. Department of Education. Faranak Rohani is the principal investigator for this research. Related information is available at http://cala.fsu.edu/ies/. Findings and opinions do not reflect the positions or policies of IES or the U.S. Department of Education.

Upgrading High-Stakes Assessments

Albert Oosterhof
Florida State University

High-stakes assessments have existed for many years. For instance, the ordinance that created the Regents examination system in New York State was passed in 1864. The use of high-stakes assessments has become widespread in recent years. In the United States, their prevalence increased with passage of the No Child Left Behind Act of 2001.

As the term "high stakes" suggests, students' performance on these assessments can result in significant actions directed at students, teachers, and/or schools. Information derived from such assessments is used summatively, not formatively.

Because of the large number of students involved, high-stakes assessments must be highly efficient with regard to administration and scoring. This constrains the formats that can be used and thus limits testing to a subset of competencies typically associated with the standards being assessed. Teachers are motivated to emphasize this subset to help their students perform well on the tests. This narrowing of the curriculum will likely become more significant as state governments consider legislation that links teachers' salaries to test scores.

By using carefully constructed performance assessments, it is possible to assess important skills that are excluded from high-stakes tests. The problem is that the cost per student to administer and score these complex assessments is high.

An Assessment Strategy

Through a three-year grant from the U.S. Department of Education,[1] the Center for Advancement of Learning and Assessment at Florida State University will examine the feasibility of a strategy that may (1) help expand the range of skills evaluated by statewide assessments and (2) add a formative aspect to these assessments.

The strategy includes three components. The first involves developing a series of performance assessments that measure selected state-level benchmarks and are administered to carefully selected samples of students. This part of the strategy estimates student proficiency at the group level and, like the National Assessment of Educational Progress (NAEP), is not designed to determine individual student proficiency. The content and number of performance

[1] The work reported in this paper is supported through a recently awarded grant from Education Research Programs at the Institute of Education Sciences (IES), award number R305A110121, administered by the U.S. Department of Education. Faranak Rohani, director of the Center for Advancement of Learning and Assessment, is the principal investigator for this research. Related information is available at http://cala.fsu.edu/ies/. Findings and opinions do not reflect the positions or policies of the IES or the U.S. Department of Education.

assessments are controlled to provide appropriate generalizability, with emphasis placed on competencies that cannot be effectively measured using conventional assessments. Using sampling would be less expensive than testing every student.

The second component focuses on the proficiency with complex skills of each individual student. It entails developing performance assessment “specifications” that define comparable measures to be developed by teachers, linking teachers’ assessments to those administered statewide to samples of students. Students’ performance levels on the teachers’ summative assessments are then compared to performance on the assessments administered statewide to samples of students. Assessments administered to these samples substantiate the reasonableness of student outcomes observed through the assessments of teachers, and vice versa.

To be effective, these expanded assessment procedures must facilitate learning rather than increase the burdens placed on teachers and school administrators. Therefore, the third component involves using the performance assessments not just summatively, but also formatively as an integral part of instruction. Results on performance-based measures would provide the basis for formative feedback to students throughout the school year.

A performance assessment presents a task for students to complete, although each specification that is used to define comparable measures will allow the generation of a family of comparable performance assessments, each representing a task relevant to the benchmark being assessed. When teachers use these performance assessments formatively during instruction, feedback to students would reference the task associated with the assessment. Feedback also would link performance on the task to the broader set of tasks implicit in the performance assessment specification. Teachers’ performance assessments eventually have a summative role, establishing what individual students have learned. However, at the classroom and school levels, the major focus is on their formative role. Emphasis is given to using meaningful feedback to guide subsequent instruction and learning. We will employ what research has identified as “best practices” related to the use of formative feedback to students.

Our research will simulate the above components in a controlled setting. We will produce and administer the external assessments that ultimately would be developed by a state assessment office. These assessments will be administered to students enrolled in participating classrooms, not statewide to samples of students. We know a NAEP-type approach can estimate achievement of groups of students from samples. We do not know whether separate external- and teacher-developed performance assessments, based on common specifications, can validate each other and help substantiate teachers’ summative assessments of individual students.

Our research occurs in the context of science instruction taught at the middle school level. The intent, however, is to establish procedures that are useful at other grade levels and in other subject areas. To maximize the benefits of the

research, we believe teachers and school administrators must assume ownership of the ideas and procedures that evolve. Therefore, a partnering relationship among teachers, curriculum specialists, measurement experts, and training specialists will be emphasized.

Knowledge and Complexity Levels

When using authentic performance assessments, it is tempting to assume that we are directly measuring the intended outcomes of learning. Any assessment, however, involves measuring something that cannot be seen. Knowledge and a student's cognitive processes are not directly observable. The performance of a student is only an *indicator* of what the student knows and is thinking.

Different types of performance are best used as indicators of different categories of knowledge. Categories often used in cognitive psychology are declarative knowledge, procedural knowledge, and problem solving. Declarative knowledge entails being able to explain things, such as how distillation works. Procedural knowledge is being able to invoke learned techniques, often in new applications, such as using distillation to separate compounds. Problem solving is involved when one has a goal but has not yet identified a means for reaching that goal. Problem solving uses strategies that rely on declarative and procedural knowledge, such as recognizing if distillation can help identify the presence of a particular chemical in a solution.

Knowledge exists at different levels of complexity. We are particularly interested in complex skills that cannot be assessed with conventional tests. For instance, a performance assessment involving procedural knowledge might ask a student to accurately determine true north, east, south, and west using sun shadows, without reference to other objects that indicate direction. A conventional test would involve a less complex task, as illustrated with this multiple-choice item:

> A stake has been placed straight up, on flat ground, in sunlight. When does the stake's shadow fall in a true north and south direction?
>
> A. When the shadow is at its shortest length (correct)
> B. When the end of the shadow is moving the fastest (incorrect)

With declarative knowledge, a performance assessment might ask a student to explain in writing (1) what you did to find true north, east, south, and west and (2) why your technique works. A less complex declarative task would be presented with this multiple-choice item:

> At any location in the United States, when does "local noon" occur?
>
> A. When the sun is exactly south of that location (correct)
> B. When local time switches from morning to afternoon (incorrect)

We have found it easier to address complexity levels after first identifying the category of knowledge (declarative, procedural, or problem solving). We do this

by using a capabilities-complexity model, in which the performance types for various levels of complexity within a given knowledge type are specified.

The complexity and difficulty of a task are distinct. Purchasing a book online uses complex procedural knowledge involving abstractions and multiple steps, but most people find it an easy task. However, most people cannot name the top-selling book of all time, although this represents low-complexity declarative knowledge. Of the three categories, the nature of declarative knowledge particularly at high complexity levels has been the least researched.

Important Issues

Our research pertaining to upgrading high-stakes assessments is just beginning. Some important issues exist that we and other researchers working in this area must address.

Any written test or performance assessment involves only a sample of tasks that might have been used for the assessment. An important question is whether the same conclusions regarding student achievement would have been reached had different, equally appropriate tasks been used. If students' performance does not "generalize," the assessment is of limited value because conclusions based on performance depend on what task was sampled. Generalizability of high-stakes assessments must be high and involves facets other than tasks (e.g., samples of raters used to score performance).

The validity of an assessment is a critical issue. Assessments always involve indicators—not direct observation—of knowledge. Validation entails establishing a link between the knowledge we seek to assess and the tasks we ask students to complete. Similarly, interpretation of performance requires establishing this link in the opposite direction, between what students were observed doing and the knowledge being assessed. Evidence-centered design will provide a framework for accomplishing this task.

Scalability and practicality are important issues that must be resolved if we are to successfully implement this alternative approach to high-stakes assessment programs.

Further Reading

Chi MTH, & Ohlsson S, (2005), Complex Declarative Learning. In KJ Holyoak & RG Morrison (Editors), *Cambridge Handbook of Thinking and Reasoning* (pp. 371–399). New York: Cambridge University Press.

Mislevy RJ, Almond RG, & Lukas JF, (2004), *A Brief Introduction to Evidence-Centered Design*, (CSE Report 632). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing. Also available at http://www.cse.ucla.edu/products/reports/r632.pdf

Oosterhof A, Rohani F, Sanfilippo C, Stillwell P, & Hawkins K, (2008), *The Capabilities-Complexity Model.* Symposium paper presented at the National Conference on Student Assessment, Orlando, FL.
Also available at http://www.cala.fsu.edu/files/ccm.pdf

Solano-Flores G, Shavelson RJ, Ruiz-Primo MA, Schultz SE, & Wiley EW, (1997), *On the Development and Scoring of Classification and Observation Science Performance Assessments*, (CSE Technical Report 458). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
Also available at http://www.cse.ucla.edu/products/Reports/TECH458.pdf

About the Author

Albert Oosterhof is professor emeritus in Educational Psychology and Learning Systems at Florida State University and is a research associate at the university's Center for Advancement of Learning and Assessment. The focus of his work is on student assessment. (albert.oosterhof@fsu.edu)