Developing Assessments of Conceptual Understanding Using "Big Ideas"

Terry P. Vendlinski, Joan L. Herman, Sam Nagashima, and Eva L. Baker
University of California, Los Angeles (CRESST)

Abstract - We describe a collaborative effort to create valid and reliable benchmark assessments for elementary school science. Classroom teachers, district administrators, professional developers, and assessment researchers designed assessments around a mapping of key ideas in each of four domains and the level of cognitive demand required to adequately assess the concept of interest. With few exceptions, this collaboration produced assessments and scoring rubrics of unusually high technical quality for locally developed assessments.

Overview

Historically, locally developed assessments, such as those developed by educators or department / district administrators, have demonstrated poor reliability and have led to inferences of questionable validity. Moreover, in the US, such assessments have often focused exclusively on the recall of facts rather than the understanding and application of key concepts in a specific domain of knowledge (McMorris & Boothroyd, 1993). Assessments created by textbook publishers are likely to suffer from the same shortcomings (Frisbie, Miranda, & Baker, 1993).

Although the intended uses of assessment data will, to some degree, dictate the kinds of evidence and hence the technical quality necessary to inform the intended decision, care must be taken to assure adequate technical quality (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). While some argue that assessments used for formative purposes need not meet the same standard of validity and reliability as high-stakes summative tests (Hubba and Freed, 2000), others suggest that accurate inferences about student understanding are necessary in formative decisions too (Black & Wiliam, 2004). It is important teachers have valid and reliable data in order to determine what to teach next, when to move on to the next topic, or when and how to best re-teach a concept. Such technical quality considerations become even more important when the data from assessments serve multiple purposes, such as when an assessment generates data that will both support formative decisions and contribute to a summative decision.

In a broad sense, validity evidence should consider the content an assessment will cover, the criterion that will be used to judge an outcome and the constructs or traits one wishes to measure. Closely associated with validity is the notion that an assessment should provide consistent (reliable) results when administered repeatedly to the same population of test takers.

While there is no single preferable approach to measuring validity or reliability, test makers are required to demonstrate that the scores produced by the assessment (and scoring instructions) are sufficiently trustworthy to justify anticipated uses and interpretations. (AERA, APA, & NCME, 1999).

Locally developed assessments that are aligned with major instructional goals, that accurately predict student performance on high-stakes assessments, that allow precise estimates of current understanding, and that correctly identify current misconceptions should be useful in making instruction both more effective and more efficient. What seems to be missing in the development of both instruction and assessment, however, are key organizing principles that function to limit the number concepts to an essential few and that logically organize topics (Schmidt, 2003).

We and colleagues at the Center for the Assessment and Evaluation of Student Learning (CAESL) have worked to bring assessments that are both aligned to goals and that focus on key concepts to reality. The CAESL framework guiding the work broadly communicates a reflective teaching process that starts with significant goals for student learning, continually assesses student understanding relative to those goals, and uses the results to guide and support student progress (DiRanna, In Press). The use of quality assessments and effective use of results must be intentionally aligned with goals for student learning (J. Herman & Baker, 2005; J. L. Herman, Osmundson, Ayala, Schneider, & Timms, 2005)

Methods

The specific project described here involved 12 teachers and district administrators from five California school districts with four professional development leaders who are also experts in science content and pedagogy and a team of three assessment researchers. The teachers and administrators involved in the test development previously had been a part of a three-year professional development program led by the participating professional developers

The group was divided into three test development teams, each charged with the development of multiple-choice, open-ended and performance assessment items for one of three intended benchmark tests: Ecosystems, Properties of Matter, and the Water Cycle. The test development process involved initial item development, expert content review, piloting and revision of items (and their scoring), selection and field testing of final test forms, and conduct of final technical analyses.

Working in small groups, composed of one professional development specialist, four educators (including district staff), and an assessment research specialist, the teams first developed the curricular flow of big ideas and subtopics in their particular topic areas. The curricular flows were based on state standards, reviews of common curriculum materials, and the participants own content and pedagogical knowledge. Teams designated concepts essential for testing, common misconceptions regarding those concepts, and identified the cognitive level at which students should be able to think about each concept — factual recall, conceptual understanding or problem solving. Generally, the understanding level was given priority.

Participants were encouraged to develop multiple-choice and open-ended items using the Assessment Design and Delivery System (ADDS). In particular, the ADDS guides designers to specify key attributes of the item such as the grade level(s) and linguistic ability the item is appropriate for, the subject area or domain of the test item, the appropriate standard being assessed, the big ideas or topics from the domain being tested, and the level of cognitive demand the item requires of the test taker. These levels—transfer, explain, complex problem solving, make connections, application, and recall—follow the cognitive demand levels developed at CRESST and are research based (see for example Baker, 1998). While item format is often associated with cognitive level (e.g., multiple choice is equated to recall), we and others encourage developers not to make that association (Stiggins, 1994). Rather, ADDS asks item developers to first consider the instructional goal, the use, and the cognitive demand of the required item, then to choose a format that best meets these needs.

Results

Based on the analysis of the data from our pilot of the multiple-choice and open-ended items, we found that CAESL-trained teachers had little difficulty collaborating to develop test items of high technical quality covering the knowledge domains in which they had teaching experience. In the case of our participants, this teaching experience did not necessarily equate to their current teaching assignment. Of the 140 multiple-choice items developed, only 5 (less than 4%) did not meet minimum technical quality specifications (1-pl model, etc.). Of these 5 items, 2 could easily have been modified to meet specifications. Half of the open-ended items were ultimately field tested and various analyses suggested that differences in student characteristics and student item interaction (rather than rubrics or raters) accounted for the vast majority of variability in test scores. Consequently, developers had little difficulty choosing items and assembling a test that both covered the curricular flow and generally met commonly accepted standards of validity and reliability.

Future Efforts

Using the benchmark assessments developed in this study, results from the fifth grade state standards test, and responses from teachers, we hope to address the question of whether student results on these tests can be used to reliably predict performance on high-stakes, state assessments.

A parallel development process, dubbed PowerSource, has also been initiated at CRESST in the domain of mathematics at the 6th through 8th grade levels. While a team of researchers is actually developing the assessment items and suggested pedagogy to teach key concepts, the development process is bases on an ontology (curricular flow) of key concepts and is similar to the development process described above.

Acknowledgements

The authors would like to thank Lynn Barakos, Diane Carnahan, Karen Cerwin, Kathy DiRanna, Craig Strang, Jo Topps, their professional development team, and all the district specialists and teachers whose efforts brought this project to fulfillment. This work is supported by the National Science Foundation under Grant Nos. ESI-0119790, and REC-0129406. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for Educational and Psychological Testing.

Baker, E. L. (1998). *Model-based performance assessment* (CSE Technical Report No. 465). Los Angeles: UCLA.

Black, P., & Wiliam, D. (2004). The formative purpose: assessment must first promote learning. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability* (Vol. 2, pp. 20 - 50).

Chicago, IL: University of Chicago Press.

DiRanna, K. (Ed.). (In Press). *Assessment Centered Teaching: A Reflective Practice*. Thousand Oaks, CA: Corwin Press.

Frisbie, D. A., Miranda, D. U., & Baker, K. K. (1993). An Evaluation of Elementary Textbook Tests as Classroom Assessment Tools. *Applied Measurement in Education, 6*(1), 21-36.

Herman, J., & Baker, E. (2005). Making Benchmark Testing Work. *Educational Leadership, 63*(3), 48 - 54.

Herman, J. L., Osmundson, E., Ayala, C., Schneider, S., & Timms, M. (2005, April). *The nature and impact of teachers' formative assessment practices.* Paper presented at the annual meeting of the American Educational Research Association (AERA), Montréal, Canada.

McMorris, R., & Boothroyd, R. A. (1993). Tests that teacher build: an analysis of clasroom tests in science and math. *Applied measurement in education, 6*(4), 321-342.

Schmidt, W. H. (2003). The Quest for a Coherent School Science Curriculum: The Need for an Organizing Principle. *Review of Policy Research, 20*(4), 569-584.

Stiggins, R. J. (1994). *Student-centered Classroom Assessment*. New York: Macmillan College Publishing.