Paper presented at the Learning Progressions in Science (LeaPS) Conference, June 2009, Iowa City, IA

UTILIZING AN ASSESSMENT DESIGN FRAMEWORK TO MAP RESEARCH QUESTIONS AND DEVELOP HIGH QUALITY ASSESSMENTS

The design of high quality science assessment items is essential to building high quality science instructional programs. But if we can't measure progress, how will we know when it occurs? Where in the process should we begin and what will guide our assessment design approaches? This paper discusses two familiar frameworks: the assessment triangle created by National Research Council (NRC 2001) and the BEAR Assessment System (Wilson 2004) and demonstrates how the use of a flexible assessment design framework not only informs the task of creating items but makes both the research design process, the design of items, and ultimately changes in instruction more accessible, coherent, and transparent.

Andrew J. Galpern, University of California, Berkeley

The NRC assessment triangle and the BEAR Assessment System

In 2001, the National Research Council published Knowing What Students Know: The Science and Design of Educational Assessment (2001).

> Any assessment is based on three interconnected elements or foundations: the aspects of achievement that are to be assessed (cognition), the tasks used to collect evidence about students' achievement (observation), and the methods used to analyze the evidence resulting from the tasks (interpretation). To understand and improve educational assessment, the principles and beliefs underlying each of these elements, as well as their interrelationships, must be made explicit. (NRC 2001)

The assessment triangle (Figure 1) provides a surprisingly simple and accessible entry point for considering the major components of the assessment-learning-teaching cycle. But exploration and elaboration of the relationships between these components is essential if these frameworks are going to be useful to researchers and teachers. In *Constructing Measures* (Wilson 2004), a much more detailed assessment design approach is presented that is organized around four building blocks (Figure 2) within a highly iterative instrument design process. One of the most promising and relevant features of Wilson's approach is its reliance on a single explicit developmental construct as the first building step in the process.

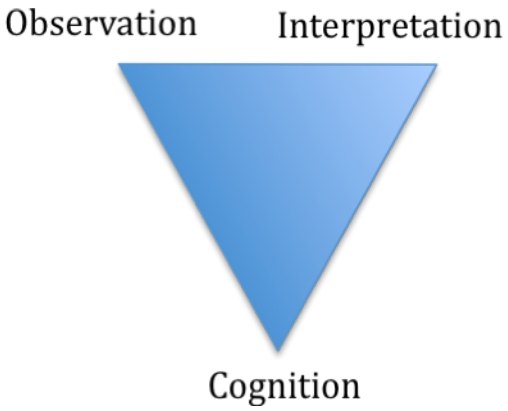

Figure 1. The National Research Council's assessment triangle. (NRC 2001)

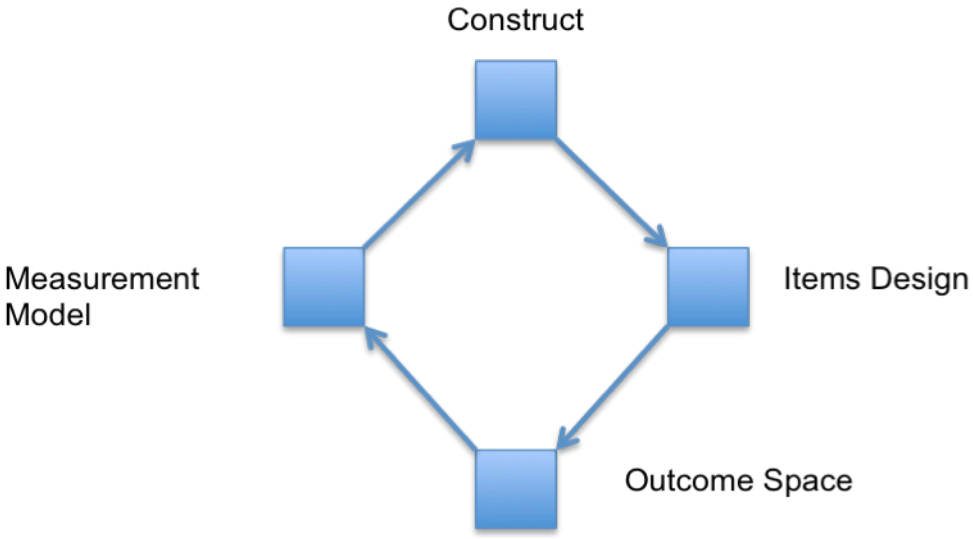

Figure 2. The four building blocks of the BEAR Assessment System (Wilson 2004)

One of the most promising and relevant features of Wilson's approach is its reliance on a single explicit developmental construct as the first building step in the process:

> A construct map, which defines a latent variable or construct, is used to represent a cognitive theory of learning consistent with a developmental perspective. This building block is grounded in the principle that assessments are to be designed with a developmental view of student learning. (Kennedy 2005)

Previous work has already shown how both the NRC and Wilson frameworks can be easily aligned with each other. This relationship can be seen in Table 1 and in Figure 3, the original diagram from Kennedy (2005).

NRC Assessment Triangle	Wilson's **Four Building Blocks**
Cognition	Construct Maps
Observation	Items Design
Interpretation	Outcome Space and Measurement Model

Table 1. The alignment between the NRC assessment triangle (NRC 2001) and the four building blocks of the BEAR Assessment System (Wilson 2004)

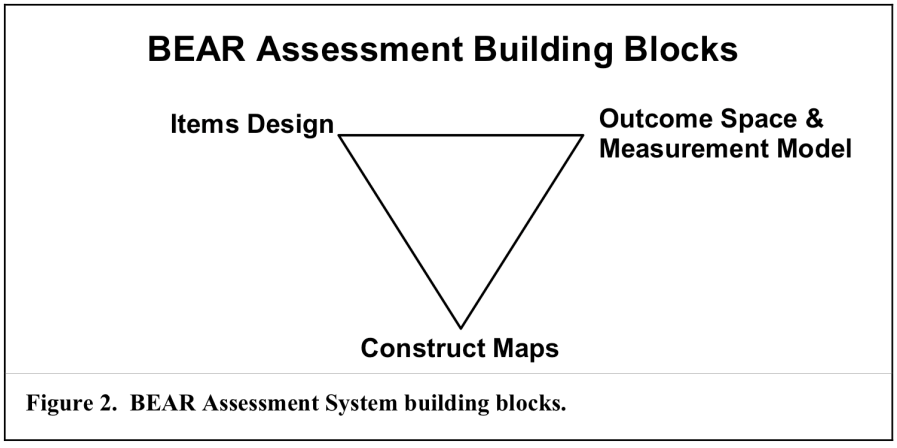

Figure 2. BEAR Assessment System building blocks.

Original diagram from The BEAR Assessment System: A Brief Summary (Kennedy 2005)

The Relationship Between Components

Although the major components of each framework are easily labeled and aligned, the need for a more detailed elaboration about the *relationship* between components (vertices) and how these relationships can inform the research design process *and* the design of items remains a rich area of research with plenty of geography to explore. In this paper, I hope to cover some of that terrain by describing several open questions in current assessment research methodology on learning progressions in science, and demonstrate how these frameworks can be used to improve the research question regarding learning progressions *as well as* the design of assessment items to locate students along those progressions.

What follows are several examples of research questions from a current project being conducted by the Berkeley Evaluation and Assessment Research Center at the University of California, Berkeley, and how these open questions map to the frameworks and help to refine the research question within the broader context of assessment, learning, and teaching.[1]

For the purposes of this demonstration and mapping, only the approach described in Wilson's *Constructing Measures* will be used, but a similar mapping could take place with the NRC assessment triangle.

Using the Framework to Map Research Questions and to Design Items

Example 1. Literacy and Language issues in Middle School Science.

The San Francisco Unified School District identified, among several priorities, the science achievement of English Language Learners and how literacy and language challenges can serve as barriers to accessing math and science content in middle school.[2] That research question can be stated simply as:

How can we overcome literacy and language challenges?

---

[1] I was suddenly struck by the earlier drafts of this paper which had these three words in a much more traditional (and much less thoughtful) order (eg. Teaching, learning, and assessment) One hope for this paper is a much more careful consideration of this mighty trinity in education. One that has the terrible habit of considering assessment at the end of the teaching and learning process.

[2] The SERP-SF project is a collaboration between the San Francisco Unified School District, the University of California at Berkeley, Stanford University, and the Strategic Education Research Partnership described here http://www.serpinstitute.org/about/field-sites/san-francisco.php in a co-development effort with middle school science teachers.

At first glance, this may appear to be a reasonable and clearly stated research question. But rather than a research question, it is actually an improperly stated goal (i.e. Goal: To overcome literacy and language challenges and improve student achievement.) Are literacy and language issues an instructional issue, a cognitive issue, or an assessment issue? The answer, unsurprisingly, is "yes". The original research question very quickly blossoms into several more specific and important research questions:

How can we study, understand, intervene, and overcome literacy and language challenges?

How can we deal with a this cluster of related research questions?

Mapping a Research Question to the Construct Map (First Building Block)

By using the BEAR Assessment System and the framework provided by the four building blocks, the need to unpack the question becomes more apparent. The first step might be to consider the potential cognitive issues regarding literacy and language. Do students with different language skills and raised in diverse cultures have a "different" way of thinking? Do students from Guatemala or China differ systematically in the way they think about matter or energy? For practical (and potentially political) reasons, the participants in this research project do not expect any kind of cognitive differences between students of different language abilities and native cultures, except to say that students who are struggling with English as a second language may also be struggling with the science concepts that language is meant to represent.

Mapping a Research Question to the Items Design (Second Building Block)

The role of the items design is to develop questions, tasks, or student "performances" that reveal student thinking along the construct (learning progression) of interest. These are typically questions or "items" on a test that take a variety of forms ranging from multiple choice questions to more open ended formats. In the case of the current project, one such learning progression is entitled "Heat Transfer Progress Map" and shown in Figure 4. Mapping the same research question from the previous section reveals a new set of design considerations and questions. How will we know English Language Learners (ELL) understand the question the way we intend? How will we ask them to respond, knowing their language skills may vary?

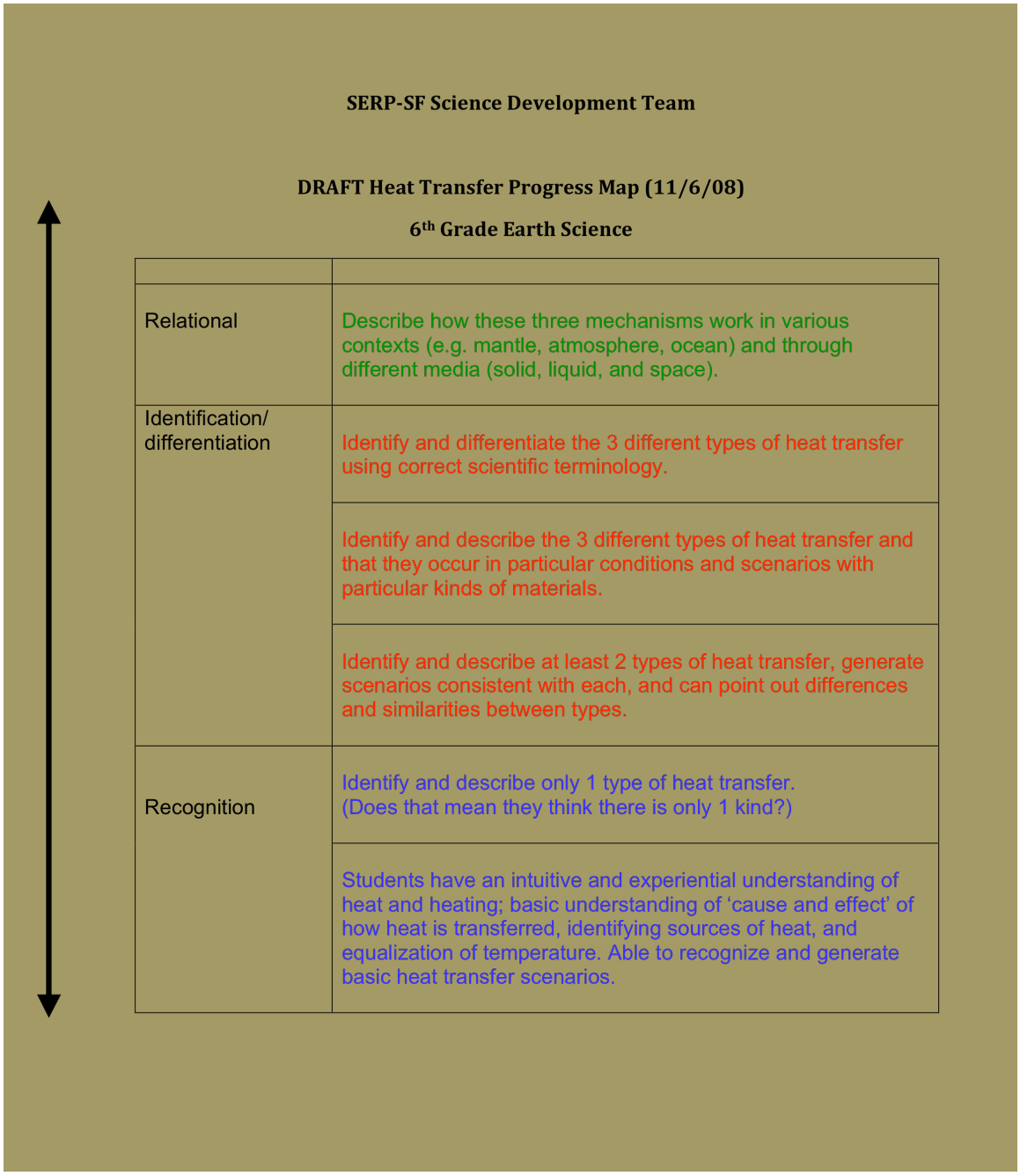

Figure 4. A screen capture of the current draft of a Heat Transfer Progress Map

For this particular project we addressed these questions in several ways. First, the teacher co-developers shared their classroom experience and specific knowledge of ELL students to describe examples of potential language issues and to develop a strategy for examining these issues. Second, they piloted items with small groups of students that included a variety of language proficiency levels. Third, they conducted "think alouds", a technique for revealing student thinking by having students talk through items to give researchers a sense of their understanding in the moment. Fourth, language in items was adjusted to reflect teacher knowledge of potentially problematic language, as well as the data collected during the think alouds. And finally, some "drawing" items were developed that had very low reading and writing burdens, intending to reveal scientific thinking less encumbered by "construct irrelevant" demands (Figure 5.)

Figure 5. Six different student responses to the drawing item "Draw and label an example for each type of heat transfer:"

Mapping a Research Question to the Outcome Space (Third Building Block)

The role of the outcome space is an organized plan to make sense of student responses in a way that reflects progress along the construct (learning progression) of interest. The outcome space typically takes the form of scoring guides or rubrics, in which every response is placed in a single category, and the categories are comprehensive and mutually exclusive. In the case of the previous drawing item shown in Figure 5, the outcome space would make sense of student drawings by noting details that revealed different levels of student thinking and ability. Mapping the same research question from the previous section reveals a new set of design considerations and questions. How will we know English Language Learners (ELL) can respond to the question the way we intend? What details are most relevant to our learning progression regarding heat transfer? Are certain heat transfer contexts more accessible to ELL students?

Mapping a Research Question to the Measurement Model (Fourth Building Block)

The role of the measurement model is to have selected an appropriate method for analyzing the data that comes out of the previous step. This analysis should include a way to report results in a format that is consistent with the original conception of our learning progression, as well as provide validity evidence for the learning progression or suggestions about how to correct it. The BEAR Assessment System is based on the idea of starting with a learning progression or construct. The measurement model it uses comes from a family of models known as Rasch models that share an approach called Item Response Theory (IRT) Because the BAS is based on a construct approach that is compatible with the IRT approach, the only measurement issues that remain are the strength of evidence for validating the learning progression and questions about the "dimensionality" of the construct. This analysis will allow us to answer more clearly whether, for example, heat transfer is a single construct or multi-dimensional.

Conclusion

The design of high quality items is essential for the measurement of learning progressions and a principled and coherent design framework such as the BEAR Assessment System is a good example of an approach that highlights the key components of the assessment-learning-teaching cycle. Although this paper limited its discussion to one small example of research design issues regarding a learning progression in middle school science, the promise of using an assessment design framework to examine and refine the research questions, as well as the design of items to test learning progressions should not be underestimated.

Current State of the Research on the Heat Transfer Learning Progression

As of the submission date of this paper (6.10.09), an initial set of student responses has been collected for the heat transfer learning progression, and we are in the process of developing and testing the scoring guides that will provide the data to be analyzed.

Future Research Plans with the Heat Transfer Learning Progression

Our future research plans include the development of additional learning progressions at the middle school science level including both content based progressions (i.e. States of Matter, Structure of the Atom, etc.) but also some more unusual work trying to outline the structure of argumentation and use of evidence in science, as a dimension, separate from specific science content.

References

Kennedy, C. A. (2005). The BEAR assessment system: A brief summary for the classroom context. BEAR Report Series, 2005-03-01. University of California, Berkeley. PDF

National Research Council (2001). Knowing What Students Know. Washington D.C.: National Academy Press. ISBN: 0309072727. A companion volume to How People Learn that reviews research on assessment of learning.

Wilson, M. (2005). Constructing measures: An item response modeling approach. Mahwah, NJ: Erlbaum.

Additional Reading[3]

Achieve. (2000). Testing: Setting the record straight. Washington, DC: Author.

American Federation of Teachers. (1999). Making standards matter 1999. Washington, DC: Author.

Appelbaum, E., Bailey, T., Berg, P., and Kalleberg, A.L. (2000). Manufacturing advantage: Why high-performance work systems pay off. Ithaca, NY: Cornell University Press.

Baker, E.L. (1997). Model-based performance assessment. Theory into Practice, 36(4), 247–254.

Barley, S.R., and Orr, J.E. (1997). Between craft and science: Technical work in U.S. settings. Ithaca, NY: Cornell University.

Baxter, G.P., and Glaser, R. (1998). Investigating the cognitive complexity of science assessments. Educational Measurement: Research and Practice, 17(3), 37–45.

Black, P., and Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education, 5(1), 7–73.

Bresnahan, T.F., Brynjolfsson, E., and Hitt, L.M. (1999). Technology, organization, and the demand for skilled labor. In M.M.Blair and T.A.Kochan (Eds.), The new relationship: Human capital in the American corporation (pp. 145–193). Washington, DC: Brookings Institution Press.

Bureau of Labor Statistics. (2000). Occupational outlook handbook, 2000–01 edition. Washington, DC: U.S. Department of Labor.

Cizek, G.J. (2000). Pockets of resistance in the education revolution. Educational Measurement: Issues and Practice, 19(1), 16–23; 33.

Cole, N.S., and Moss, P.A. (1993). Bias in test use. In R.L.Linn (Ed.), Educational measurement (Third Edition) (pp. 201–220). Phoenix, AZ: American Council on Education and The Oryx Press.

Council of Chief State School Officers. (1999). Data from the annual survey. State student assessment programs. Volume 2. Washington, DC: Author.

Dwyer, C.A. (1998). Assessment and classroom learning: Theory and practice.

[3] This reference list is included in the first chapter references of National Research Council (2001). Knowing What Students Know. Washington D.C.: National Academy Press. ISBN: 0309072727. A companion volume to How People Learn that reviews research on assessment of learning.

Assessment in Education, 5(1), 131–137.

Edelson, D.C., Gordon, D.N., and Pea, R.D. (1999). Designing scientific investigation environments for learners: Lessons from experiences with scientific visualization. Journal of the Learning Sciences, 8(3/4), 391–450.

Education Week. (1999). Quality counts '99: Rewarding results, punishing failure. Bethesda, MD: Author.

Finn, C.E., Jr., Petrilli, M.J., and Vanourek, G. (1998). The state of state standards. Fordham Report (Volume 2). Washington, DC: The Thomas B.Fordham Foundation.

Glaser, R., Linn, R., and Bohrnstedt, G. (1997). Assessment in transition: Monitoring the nation's educational progress. New York: National Academy of Education.

Glaser, R., and Silver, E. (1994). Assessment, testing, and instruction: Retrospect and prospect. In L.Darling-Hammond (Ed.), Review of research in education (Volume 20). (pp. 393–419). Washington, DC: American Educational Research Association.

Klein, S.P., Hamilton, L.S., McCaffrey, D.F., and Stecher, B.M. (2000). What do test scores in Texas tell us? Santa Monica, CA: RAND.

Koretz, D.M., and Barron, S.I. (1998). The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS). Santa Monica, CA: RAND.

Lemann, N. (1999). The big test: The secret history of the American meritocracy. New York: Farrar, Straus, and Giroux.

Lindquist, E.F. (1951). Preliminary considerations in objective test construction. In E.F.Lindquist (Ed.), Educational measurement (pp. 119–184). Washington, DC: American Council on Education.

Linn, R. (2000). Assessments and accountability. Educational Researcher, 29(2), 4–16.

Linn, R.L., Baker, E.L., and Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20(8), 15–21.

Mehrens, W.A. (1998). Consequences of assessment: What is the evidence? Educational Policy Analysis Archives, 6(13). <http://epaa.asu.edu/epaa/v6n13.html>. [March 28, 2000].

Messick, S. (1984). The psychology of educational measurement. Journal of Educational Measurement, 21(3), 215–237.

Mislevy, R.J. (1993). Foundations of a new test theory. In N.Frederiksen, R.J.Mislevy, and I.I.Bejar (Eds.), Test theory for a new generation of tests. Hillsdale, NJ: Erlbaum.

Mislevy, R.J. (1994). Test theory reconceived. (CSE Technical Report 376). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

National Academy of Education. (1996). Implications for NAEP of research on learning and cognition. Stanford, CA: Author.

National Center for Education Statistics. (1996). Technical issues in large-scale performance assessment. Washington, DC: U.S. Department of Education.

National Council of Teachers of Mathematics. (2000). Principles and standards for school mathematics. Reston, VA: Author.

National Research Council. (1996). National science education standards. National Committee on Science Education Standards and Assessment. Coordinating Council for Education. Washington, DC: National Academy Press.

National Research Council. (1999a). The changing nature of work: Implications for occupational analysis. Committee on Techniques for the Enhancement of Human Performance: Occupational Analysis. Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council. (1999b). Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress. Committee on the Evaluation of National and State Assessments of Educational Progress. J.W. Pellegrino, L.R.Jones, and K.J.Mitchell, (Eds.). Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council. (1999c). High stakes: Testing for tracking, promotion, and graduation. Committee on Appropriate Test Use. J.P.Heubert and R.M.Hauser, (Eds.). Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council. (2001). Building a workforce for the information economy. Committee on Workforce Needs in Information Technology. Board on Testing and Assessment; Board on Science, Technology, and Economic Policy; and Office of Scientific and Engineering Personnel. Washington, DC: National Academy Press.

New Standards™ . (1997). Performance standards: English language arts, mathematics, science, applied learning (Volume 1, Elementary school). Washington, DC: National Center for Education Statistics and the University of Pittsburgh.

Nichols, P.D. (1994). A framework for developing cognitively diagnostic assessments. Review of Educational Research, 64(4), 575–603.

Pellegrino, J.W., Baxter, G.P., and Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. In A. Iran-Nejad and P.D.Pearson (Eds.), Review of research in education (Volume 24) (pp. 307–353). Washington, DC: American Educational Research Association.

Popham, W.J. (2000). Modern educational measurement: Practical guidelines for educational leaders. Needham, MA: Allyn and Bacon.

Resnick, L.B., and Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R.Gifford and M.C.O'Connor (Eds.), Changing assessments: Alternative views of aptitude, achievement, and instruction. Boston: Kluwer.

Rothman, R., Slattery, J.B., Vranek, J.L., and Resnick, L.B. (in press). The alignment of standards and assessments. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing. Graduate School of Education, University of California.

Secretary's Commission on Achieving Necessary Skills (SCANS). (1991). What work requires of schools: A SCANS report for America 2000. Washington, DC: U.S. Department of Labor.

Snow, R.E., and, D.F. (1989). Implications of cognitive psychology for educational measurement. In R.L.Linn (Ed.), Educational measurement (3rd Edition), (pp. 263–330). New York: Macmillan.

Steele, C.M. (1995). Stereotype threat and the intellectual test performance of African Americans. Journal of Personality and Social Psychology, 69(5), 797–811.

Steele, C.M. (1997). How stereotypes shape intellectual identity and performance. American Psychological Association, 55(6), 613–629.

U.S. Congress, Office of Technology Assessment. (1992). Testing in American schools: Asking the right questions. Washington, DC: U.S. Government Printing Office.

Wilson, M., and Adams, R.J. (1996). Evaluating progress with alternative assessments: A model for chapter 1. In M.B.Kane (Ed.), Implementing performance assessment: Promise, problems, and challenges. Hillsdale, NJ: Lawrence Erlbaum Associates.