CRESST LINE
Newsletter of the National Center for Research on Evaluation, Standards, and Student Testing
SPRING 2004

Joan L. Herman, Eva L. Baker, Robert L. Linn

FROM THE DIRECTORS:

ACCOUNTABILITY SYSTEMS IN SUPPORT OF STUDENT LEARNING: MOVING TO THE NEXT GENERATION

Joan L. Herman, Eva L. Baker, and Robert L. Linn

Standards-based reform and the accountability systems it has generated are built on the assumption that being clear about learning goals, measuring students' progress toward them, and applying known consequences will help to improve student learning. Embedded strongly in *No Child Left Behind* ([NCLB] 2002), the logic seemingly is straightforward and linear, even if the reality is not: We agree on standards for what students ought to know and be able to do, agree as a society and as communities of educators and schools that we will work with and expect *all* students to achieve these standards, develop measures that tell us how well we (educators, students, student subgroups) are doing, and then use feedback from these measures and marshal all available resources to improve educational opportunities and to assure every student's success.

However, while conveyed as technical steps in a familiar and time-honored problem-solving process, current accountability schemes at heart are systems for motivating performance. NCLB makes this point clear by attaching severe and escalating consequences for schools that fail to meet goals for Adequate Yearly Progress (AYP). Ostensibly providing a technical system for measuring performance and providing data to support improvement, NCLB's policy objective, of course, is to stimulate state and local action to significantly increase student learning and to assure that schools make progress toward all students achieving established standards.

continued on next page

Inside

If we accept that accountability systems are intended to serve both symbolic and technical functions, we can ask how well they are operating symbolically to motivate action and how well they are doing in providing adequate technical information to support intended inferences. In this column we argue that while accountability may be accomplishing its motivational goals, current systems may be technically inadequate for serving ultimate goals of improving student learning. While the federal government recently has offered states new flexibility in counting students for AYP purposes, important issues remain in optimizing the learning value of accountability assessments. Unless and until they seriously incorporate local and/or classroom assessment (performance assessment), accountability systems cannot assure adequate reliability and validity for *individual* decision-making purposes and cannot provide necessary information to support student learning. We present one idea for system design.

…teachers and administrators listen well to the signals sent by important large-scale assessments, model what they see in the tests, and adapt their curriculum and teaching accordingly…

Symbolic Purposes

As a policy lever, accountability systems can serve symbolic purposes in a variety of ways, among them: They establish educational improvement as a vital public priority; they create specific targets for improvement efforts; they communicate to educators, administrators, and parents what is expected; they provide incentives and/or sanctions for reaching specified goals; and thereby they are intended to motivate all levels of the education system to focus on achieving the specified policy goals.

Ample research suggests that accountability systems may be a reasonably effective mechanism in serving such functions. According to the literature, teachers and administrators listen well to the signals sent by important large-scale assessments, model what they see in the tests, and adapt their curriculum and teaching accordingly, focusing their instruction and improvement efforts on areas that test results suggest are in need of attention (Borko & Elliott, 1998; Borko & Stecher, 2001; Firestone, Camilli, Yurecko, Monfils, & Mayrowetz, 2000; Firestone, Mayrowetz, & Fairman, 1998; Goldberg & Rosewell, 2000; Herman & Klein, 1996; Koretz, Barron, Mitchell, & Stecher, 1996; Koretz, McCaffrey, Klein, Bell, & Stecher 1993; Lane, Stone, Parke, Hansen, & Cerrillo, 2000; McDonnell & Choisser, 1997; Smith & Rottenberg, 1991; Stecher, Barron, Chun, & Ross, 2000; Stecher, Barron, Kaganoff, & Goodwin, 1998; Wolf & McIver, 1999). Even as this same body of research suggests the curriculum distortions that large-scale assessments can foster, it makes clear the symbolic and political power of accountability and assessment.

From Symbolic to Information Value

Symbolic and political purposes aside, however, what of the information value that accountability assessments are supposed to provide? The technical issues are many in assuring the reliability and validity of measures for specific decision-making purposes. Putting first things first, the alignment between a state's standards and its assessment is a bedrock issue for standards-based systems, yet studies of state tests show uneven results (see, for example, Bhola, Impara, & Buckendahl, 2003; Herman, Webb, & Zuniga, 2002; Rothman, Slattery, Vranek, & Resnick, 2002; Webb, 1999). Existing instruments, it appears, have not been systematically designed to optimize concurrence, comprehensiveness, and balance relative to state standards. Yet to the extent that state assessments give short shrift to some standards, as the research cited above suggests, the instruction and teaching of children are apt to do likewise.

The case for multiple measures. That a single test cannot address all that is important for students to know and be able to do is axiomatic. Multiple measures are needed to address the full depth and breadth of our expectations for student learning. Further, the multiple-choice and short-answer type items that tend to predominate in large-scale accountability tests can go only so far in tapping the complex thinking, communication, and problem-solving skills that students will need for future success. Other types of performance measures—essays, applied projects, portfolios, demonstrations, oral presentations, etc.—are needed to represent and guide students' progress. Moreover, multiple types of measures can better respond to the reality of individual differences than can a single test: Just as not all students learn in the same way, not all students can demonstrate their proficiency in the same way. Some may do better in some formats and contexts than in others. If important decisions are based on results, students need multiple avenues for showing what they know and can do. Professional testing standards are clear on this issue: A single test should never be used as the sole determinant of any important decision.

continued on page 4

Don't Join the Herd

Attend the 2004 CRESST Conference

September 9–10
University of California, Los Angeles

Tired of crowded conferences that answer few, if any, of your most important questions?

The CRESST conference is a unique, two-day experience on the beautiful UCLA campus, featuring presentations by many of the top accountability experts in the nation.

On September 9–10, 2004, more than 40 conference presenters will focus on urgent accountability topics to help answer many questions posed by today's increasingly complex assessment environment. Attending the CRESST conference connects you to a research organization that treats you like part of the family, not part of a herd.

When the conference is over, we will be here to help.

Just a few of our confirmed presenters include:

The CRESST Co-directors, Eva Baker, Robert Linn, and Joan Herman
Michael Cohen, ACHIEVE
Michael Kirst, Stanford University
Hilda Borko, University of Colorado at Boulder
Robert Glaser, University of Pittsburgh
Daniel Koretz, Harvard University
Richard Shavelson, Stanford University
Jamal Abedi, CRESST/UCLA
Robert Mislevy, University of Maryland
Edward Haertel, Stanford University
Stephen Dunbar, University of Iowa/Iowa Tests of Basic Skills
Geno Flores, California Department of Education
Linda Darling-Hammond, Stanford University
Scott Marion, The Center for Assessment
Harold F. O'Neil, University of Southern California
Brian Stecher, RAND
Noreen Webb, CRESST/UCLA

See you in September!

More information, including a downloadable registration form, is available on the CRESST Web site, CRESST.org, or send a note to Kim Hurst at kim@cse.ucla.edu

CRESST

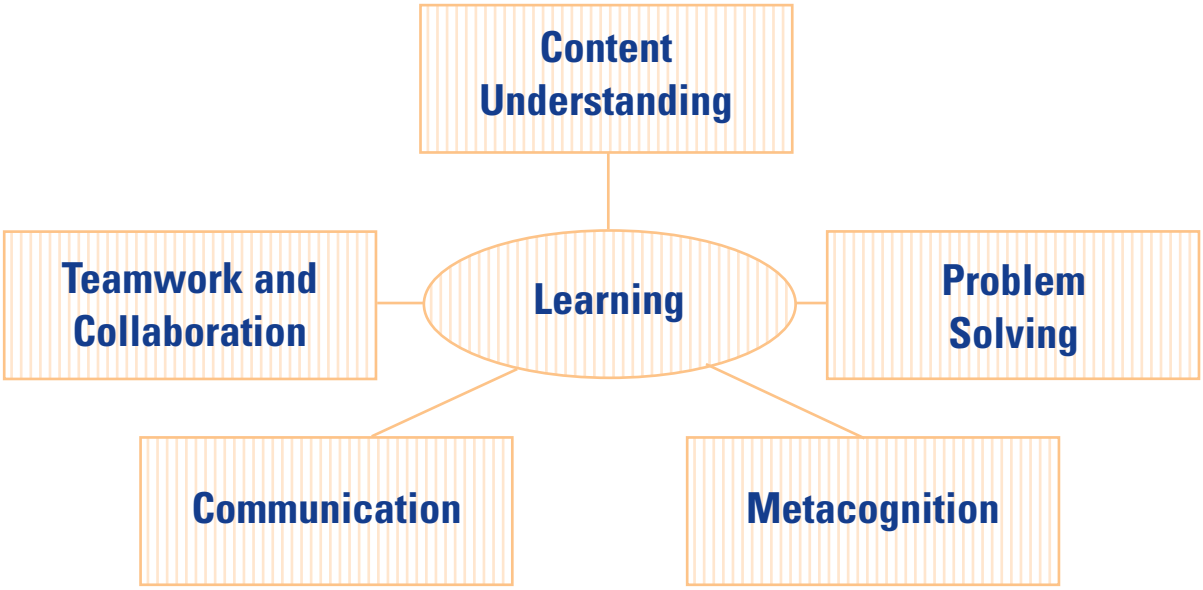

Figure 1. CRESST model-based assessment–Types of learning.

Multiple measures integrated with classroom practice. While the potential for errors in measurement provides one strong technical base for multiple measures, the need for ongoing information to guide teaching and learning provides another. Standards-based reform makes clear the importance of setting clear goals, establishing plans to achieve them, and using data to monitor progress and adjust plans accordingly, a continuous improvement cycle held in common with various reform strategies past and present in education and elsewhere (see, for example, Bloom, 1968, 1971, 1981; Cooley & Glaser, 1968; Deming, 1986; Fuhrman & Elmore, 2004; Glaser, 1966a, 1966b; Resnick & Resnick, 1992; Suppes, 1964).

No matter how well aligned and how sensitively crafted, however, once-or-twice-a-year accountability assessments can offer only a limited perspective on what children really know and can do, the nature of their thinking, the sources of their strengths and misconceptions, and the factors that may be impeding their progress. In order to understand why student performance is as it is and to get to the root of whatever teaching and learning issues may exist, schools and teachers must move to a more detailed level of assessment and analysis than annual state tests afford. Schools and teachers need to be able to supplement external assessment results with other local data, both to acquire the deep understanding they need to improve the learning process, and to get regular information during the school year about whether and how students are progressing. The once-or-twice-a-year test results arrive too late to make a difference in student learning. Rather, assessment must be ongoing, teaching must be tailored to students' specific strengths and needs, and students who are faltering need to be given special attention. Recent research indeed underscores the power of formative classroom assessment (Black & Wiliam, 1998).

Coordinated systems of assessment. District, school, and/or classroom assessments that are aligned with standards and coordinated with accountability tests are needed to provide educators with the diverse and regular forms of evidence that they must have to understand and improve their students' learning. Results from such measures can afford a comprehensive picture of students' achievement relative to expected standards and, potentially in combination with state or accountability tests, can be configured to provide a strong and rich base of evidence for decisions about students' proficiencies.

In arguing for coordinated systems of assessments, we realize the potential burden and the need for flexibility and local control. A single system of specific assessments is unlikely to be sensitive to the variations across districts, schools, and classrooms in particular curricula, individual students' needs and cultures, local preferences for assessment types, and topical priorities. Any workable system must permit local choices by teachers, schools, and districts with regard to when, how, and the specifics of what to assess. However, at the same time, to be useful for accountability purposes, a system

must provide some degree of comparability of results. We believe that model-based assessment systems offer the potential for such a system.

Using the CRESST Model-Based Assessments as a Basis for a Flexible, Integrated System

The CRESST model-based assessments are based on the assumption that there are core types of learning outcomes that recur in every area of curriculum, and that these, deeply embedded in disciplinary content, offer appropriate foci for instruction and assessment across curriculum areas. As Figure 1 shows, these learning types include content understanding, problem solving, communication, teamwork and collaboration, and metacognition (Baker, 2003).

The model for each of these learning types represents a distillation of scientific findings about learning for each family of cognition. Each model then is made concrete in one or more templates that represent standard and replicable assessment approaches that can be used for both large-scale and classroom assessment purposes. For example, CRESST has synthesized from research a model of how students display content understanding (Baker, forthcoming), emphasizing principled structure and use of prior knowledge as major components. Two different templates have been developed, one using writing explanations given primary source materials, and the other using computer-based knowledge mapping to display comprehension (Baker, 1994; Chung, O'Neil, & Herl, 1999). Both templates rely on expert approaches to scoring. In problem solving, the models have emphasized the use of problem identification strategies, prior knowledge, and metacognitive strategies. Templates have been developed that emphasize the process of problem solving, explanation, and collaborative performance. Once again, scoring rubrics are based on expert performance. Detailed specifications define each template and embed any assessment in the appropriate content to be assessed. Task specifications also describe the kinds of information and content material to be provided for students, the logistics of time and arrangements, and training required for judging student responses validly (Baker, Aschbacher, Niemi, & Sato, 1992). The specifications provide a general blueprint defining what is expected of students and delimiting the nature of the assessment.

These templates or blueprints then can be instantiated in disciplinary content, reflecting the domain-specific knowledge and skills that are expected of students, and used across any number of specific topics, as CRESST has done in history, mathematics, literature, and science (see, for example, Baker, 2004; Goldschmidt & Martinez-Fernandez, 2004; Niemi & Baker, 1998; Niemi, Chen, & Steinberg, 2004; Niemi, Sylvester, & Baker, 1998; Waltman & Baker, 1997; Wang & Wang, 2004). The idea is not that the CRESST models can or do represent all types of or all possible approaches to the assessment of learning, but rather that they can be developed by states or locales to represent a reasonable set of expectations that can serve to focus instruction and clearly communicate expectations to all relevant stakeholders—teachers, students, parents, etc., even test developers. Everyone knows what to expect, in terms of the desired content (knowledge and skills), anticipated cognitive demands, and other characteristics of tasks or

- Overall Content Understanding—A holistic score related to the appropriateness of the essay and the quality of understanding displayed.
- Use of Principles—Incorporates important and appropriate principles as organizing ideas.
- Use of Prior Knowledge—References prior knowledge as evidence to explain ideas or to justify analysis.
- Use of Text—Incorporates text materials included with the assessment to illustrate or extend argument.
- Integrated Argument—Uses an integrative system of argument that is appropriate to the subject area.
- Misconceptions—Number and type.

Figure 2. CRESST rubric for assessing conceptual understanding.

items in which students will be engaged and the nature and quality of expected student performance. Yet the specifics of those expectations can be embedded in a wide variety of materials and topics. And they can support transfer of skills to new domains.

That the CRESST models are framed in terms of core types of learning that cross curriculum areas also means that they can form an intermediate structure to partially manage the alignment of standards, instruction, and assessment (Baker, in press). That is, regardless of subject area, it is possible to classify a state's content standards with regard to whether they call on basic facts, conceptual understanding, problem solving, communication, etc. Having once determined the type(s) of learning a standard calls for, suitable blueprints can be accessed and appropriate assessments developed. The blueprints and assessments could be customized to more specific skills for those states whose standards warrant it.

CRESST model-based assessments represent one approach to developing flexible and technically sound classroom and local assessments that are purposively linked to state standards and to statewide standards-based tests.

Blueprints or templates, including type of content and scoring rubrics, could be used for classroom assignments and assessments (Baker, 2004), as well as for statewide or local assessments. The CRESST model of content understanding uses an analytic rubric since research shows that the rubric dimensions provide a productive focus for instruction. CRESST's generalized rubric for assessing conceptual understanding (Figure 2), which is derived from studies examining the differences between experts' and novices' responses (see, for instance, Baker, Freeman, & Clayton, 1991; Chi & Glaser, 1980; Chi, Glaser, & Rees, 1982; Larkin, McDermott, Simon, & Simon, 1980; Newell & Simon, 1972), provides an example.

Consider how such a rubric, or selected dimensions from it, could be used to assess students' understanding in a variety of tasks: a 30-minute task that asks students to explain why particular objects sink or float; or a week-long project in which students explain what will happen to moving objects of various sizes as they come into contact with a new force; or a semester-long project to design a floating armada. Common sets of criteria can be used to address tasks at various levels of complexity and sophistication, providing a consistent tool for scaffolding students' growing understandings. At the same time, through careful task specification, tasks can be systematically designed to tap various levels of content complexity, meaning also that comparable tasks can be developed in the context of various specific topics of interest. The intended system thus is flexible in permitting variation but still can yield the types of comparable data that are needed for large-scale assessment results.

Building Capability for More Productive Assessment Systems for the Future

CRESST model-based assessments represent one approach to developing flexible and technically sound classroom and local assessments that are purposively linked to state standards and to statewide standards-based tests. This solution moves beyond expecting local educators to develop their own assessments—a time-intensive task for which their background and training offers little sound preparation—and instead provides educators with customizable tools that they can use to actually implement standards-based learning. A key consideration is that the models have elements that can be adapted and reused for different grade levels and topics, providing one ingredient toward a coherent system. Reusable templates and rubrics reduce assessment development costs, support instructional planning and feedback, and provide a way to link external testing with classroom testing.

We recognize the many complexities of moving forward to such a system, including issues of design, delivery, and comparability, and strategies for assuring both the technical quality of the system and that teachers have the capacity to use assessment to support their students' learning. We think that technology may hold part of the solution—for example, as a platform for efficient delivery, training, and scoring. The collective knowledge of the field across psychometrics, assessment, cognitive theory, and teaching and learning holds another part of the answer, and all must be leavened with the sociopolitical realities and challenges of instituting change. We feel the prospects are exciting and will be moving forward with them.

References

Baker, E. L. (1994). Learning-based assessments of history understanding. *Educational Psychologist, 29,* 97-106.

Baker, E. L. (2003). Multiple measures: Toward tiered systems. *Educational Measurement: Issues and Practice, 22*(2), 13-17.

Baker, E. L. (2004, April). From research to impact: Scaling up model-based assessments. In J. Evans (Chair), *Applying research-based performance assessment models in routine practice in a large urban school district: The pleasure-pain principle.* Symposium presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Baker, E. L. (in press). Aligning curriculum, standards, and assessments: Fulfilling the promise of school reform. In C. A. Dwyer (Ed.), *Measurement and research in the accountability era*. Mahwah, NJ: Lawrence Erlbaum Associates.

Baker, E. L. (forthcoming). Principles for scaling up: Choosing, measuring effects, and promoting the widespread use of educational innovation. In B. Schneider (Ed.), *Proceedings of the Data Research and Development Center Conference "Conceptualizing scale-up: Multidisciplinary perspectives."* Chicago, IL: University of Chicago, NORC.

Baker, E. L., Aschbacher, P. R., Niemi, D., & Sato, E. (1992). *CRESST performance assessment models: Assessing content area explanations*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Baker, E. L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp.131-153). Englewood Cliffs, NJ: Prentice-Hall.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice, 22*(3), 21-29.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5,* 7-74.

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment, 1*(2), 1-5.

Bloom, B. S. (1971). Mastery learning. In J. H. Block (Ed.), *Mastery learning: Theory and practice* (pp. 47-63). New York: Holt, Rinehart and Winston.

Bloom, B. S. (1981). *All our children learning*. New York: McGraw-Hill.

Borko, H., & Elliott, R. (1998). *Tensions between competing pedagogical and accountability commitments for exemplary teachers of mathematics in Kentucky* (CSE Tech. Rep. No. 495). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Borko, H., & Stecher, B. M. (2001, April). Looking at reform through different methodological lenses: Survey and case studies of the Washington state education reform. In J. Manise (Chair), *Testing policy and teaching practice: A multi-method examination of two states*. Symposium presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Chi, M. T. H., & Glaser, R. (1980). The measurement of expertise: Analysis of the development of knowledge and skill as a basis for assessing achievement. In E. L. Baker & E. S. Quellmalz (Eds.), *Educational testing and evaluation: Design, analysis, and policy* (pp. 37-47). Beverly Hills, CA: Sage Publications.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 1, pp. 7-75). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chung, G. K. W. K., O'Neil, H. F., Jr., & Herl, H. E. (1999). The use of computer-based collaborative knowledge mapping to measure team processes and team outcomes. *Computers in Human Behavior, 15,* 463-493.

Cooley, W. W., & Glaser, R. (1968). *An information and management system for individually prescribed instruction*. Pittsburgh, PA: Pittsburgh University, Learning Research and Development Center.

Deming, W. E. (1986). *Out of the crisis. Quality, productivity and competitive position*. Cambridge: Cambridge University Press.

Firestone, W. A., Camilli, G., Yurecko, M., Monfils, L., & Mayrowetz, D. (2000, April). *State standards, socio-fiscal context and opportunity to learn in New Jersey*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. [Also available in *Educational Policy Analysis Archives 8*(35). Retrieved August 23, 2002, from http://epaa.asu.edu/epaa/v8n35/]

Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis, 20,* 95-113.

Fuhrman, S. H., & Elmore, R. F. (Eds.). (2004). *Redesigning accountability systems for education*. New York: Teachers College Press.

Glaser, R. (1966a). *The program for individually prescribed instruction*. Pittsburgh, PA: Pittsburgh University, Learning Research and Development Center.

Glaser, R. (1966b). Studies of the use of programmed instruction in the intact classroom. *Psychology in the Schools, 3,* 318-333.

Goldberg, G. L., & Rosewell, B. S. (2000). From perception to practice: The impact of teachers' scoring experience on performance-based instruction and classroom practice. *Educational Assessment 6,* 257-290.

Goldschmidt, P. G., & Martinez-Fernandez, J.-F. (2004, April). Teachers' dual roles: Providing opportunity to learn and judging the outcomes. In J. Evans (Chair), *Applying research-based performance assessment models in routine practice in a large urban school district: The pleasure-pain principle*. Symposium presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Herman, J. L., & Klein, D. (1996). Evaluating equity in alternative assessment: An illustration of opportunity to learn issues. *Journal of Educational Research 89,* 246-256.

Herman, J. L., Webb, N., & Zuniga, S. (2002, April). *Alignment and college admissions: The match of expectations, assessments, and educator perspectives*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Koretz, D., Barron, S., Mitchell, K., & Stecher, B. (1996). *The perceived effects of the Kentucky Instructional Results Information System (KIRIS)* (MR-792-PCT/FF). Santa Monica, CA: RAND.

Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1993). *The reliability of scores from the 1992 Vermont Portfolio Assessment Program* (CSE Tech. Rep. No. 355). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Lane, S., Stone, C. A., Parke, C. S., Hansen, M. A., & Cerrillo, T. L. (2000, April). *Consequential evidence for MSPAP from the teacher, principal and student perspective*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Larkin, J. H., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science, 208,* 1335-1342.

McDonnell, L. M., & Choisser, C. (1997). *Testing and teaching: Local implementation of new state assessments* (CSE Tech. Rep. No. 442). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Newell, A., & Simon, H. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Niemi, D., & Baker, E. L. (1998, January). *Design and development of a comprehensive assessment system: Pilot testing, scoring, and refinement of mathematics and language arts performance assessments* (Final Deliverable). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Niemi, D., Chen, E. J., & Steinberg, D. H. (2004, April). Validating a large-scale performance assessment development effort. In J. Evans (Chair), *Applying research-based performance assessment models in routine practice in a large urban school district: The pleasure-pain principle*. Symposium presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Niemi, D., Sylvester, R. M., & Baker, E. L. (1998, November). *Design and development of a comprehensive assessment system: Identification and pilot testing of performance assessments and validity studies development* (Final Deliverable). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 35-75). Boston: Kluwer.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *The alignment of standards and assessments* (CSE Tech. Rep. No. 566). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Smith, M. L., & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice, 10*(4), 7-11.

Stecher, B. M., Barron, S. I., Chun, T., & Ross, K. (2000). *The effects of Washington education reform on schools and classrooms* (CSE Tech. Rep. No. 525). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Stecher, B. M., Barron, S. I., Kaganoff, T., & Goodwin, J. (1998). *The effects of standards-based assessment on classroom practices: Results of the 1996-97 RAND survey of Kentucky teachers of mathematics and writing* (CSE Tech. Rep. No. 482). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Suppes, P. (1964). Modern learning theory and the elementary-school curriculum. *American Educational Research Journal, 1, 79-93.*

Waltman, K. K., & Baker, E. L. (1997, September). *Design and development of a comprehensive assessment system: Summary of the 1996-1997 large-scale pilot tests, scoring sessions, and teacher feedback* (Final Deliverable). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Wang, J., & Wang, H. (2004, April). Predictive validity and relative fairness of writing performance assessments. In J. Evans (Chair), *Applying research-based performance assessment models in routine practice in a large urban school district: The pleasure-pain principle*. Symposium presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Webb, N. L. (1999). *Research Monograph No. 18: Alignment of science and mathematics standards and assessments in four states*. Madison: University of Wisconsin, National Institute for Science Education.

Wolf, S. A., & McIver, M. C. (1999). When progress becomes policy: The paradox of Kentucky state reform for exemplary teachers. *Phi Delta Kappan, 80,* 401-406.