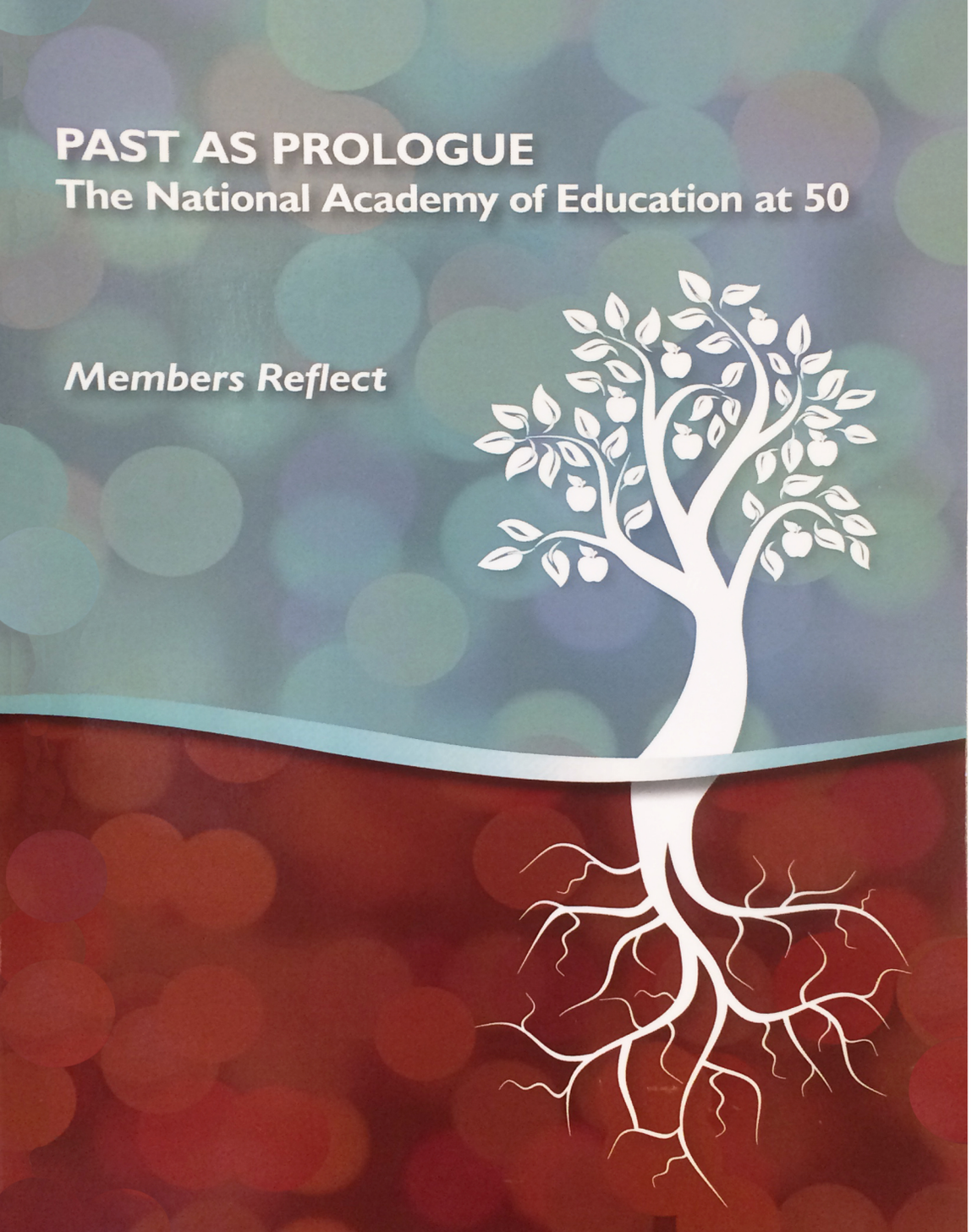

Feuer, M.J., Berman, A.I., & Atkinson, R.C. (Eds). (2015). *Past as Prologue: The National Academy of Education at 50. Members Reflect.* Washington, DC: National Academy of Education

If We Know So Much from Research on Learning, Why Are Educational Reforms Not Successful?

Lorrie A. Shepard[1]

In this essay, I return to the question that framed my presidential address to the National Academy of Education (NAEd) in 2009. Given a deep research base on learning, why are educational reforms not successful? At that time, a new President Obama was beginning his first term in office with great hope, and during the preceding election year the NAEd had provided to both political parties a series of white papers summarizing research relevant to key education issues. One of those white papers, *Standards, Assessments, and Accountability* (Shepard, Hannaway, & Baker, 2009), was the focus of my address titled "Curricular Incoherence: The Story of Educational Reforms Undone."

Back in 2009, my intention had been to explain the bait-and-switch errors that arose in previous reforms when similarly named policies were substituted for research-inspired ideals. My hope was to forestall the problems that arise from superficial understandings. Now with 6 years elapsed, we see disappointingly that cautions—issued by many—were not heeded, and ill effects from top-down accountability mandates continue to pervade the education landscape. To understand why these patterns continue and what might be done about it, I repeat again the old history of standards-based reforms in the 1990s, attending in particular to the connections that reformers drew then to advances in cognitive

[1] Lorrie A. Shepard is the Dean and Distinguished Professor of the School of Education, University of Colorado Boulder. She was elected to the National Academy of Education in 1992.

science and research on learning. I then explain how reforms have been undone by direct attacks, but more often by competing visions; or they are subverted by superficial understandings and lack of support. In the last section of the essay, I repeat the most important of the NAEd white paper recommendations regarding standards and assessments. In the case of assessment reforms, I try to explain why short-cut versions of imagined reforms have again fallen short of what might have been possible; I also consider which if any of the best ideas in that paper might still be worth pursuing. I conclude with a plea to roll back accountability mandates, which have only exacerbated inequities, and to invest instead in smaller-scale reforms designed to support teaching and learning.

The term "standards-based reforms" was coined in the 1990s to signify the central role that "world-class standards" were expected to play both in raising expectations and in crafting coherent systems to meet these higher standards. A spate of policy reports condemned the existing, de facto basic skills curriculum driven by standardized tests and textbooks and called for the creation of challenging standards aimed at higher-order thinking and problem-solving abilities. As part of an aligned system, new forms of authentic assessments requiring more fulsome enactments of ambitious learning goals were expected to leverage reform efforts instead of misdirecting teaching and learning as previous tests had done.

Leading education researchers were deeply involved alongside policymakers in arguing for standards and accompanying reforms. Mathematics educators led other disciplines in developing curriculum and evaluation standards that sought to change the character of what mathematics was thought to be as well as how it was taught (National Council of Teachers of Mathematics, 1989; National Research Council, 1989). Mike Smith and Jennifer O'Day (1990) wrote an iconic piece describing their vision of "systemic school reform," which in contrast to local school restructuring reforms would build out new, content-driven state structures. Clear and challenging standards for student learning would provide an organizing framework toward which other policy tools could be focused. Lauren Resnick sat on a dozen policy commissions and taught politicians and business leaders about the cognitive science behind the "thinking curriculum." She also helped explain why the decomposability and decontextualization assumptions implicit in the machinery of standardized tests were inimical to teaching for deep understanding (Resnick & Resnick, 1992).

Among many policy documents, a report of the NAEd (McLaughlin & Shepard, 1995) focused in particular on the learning theory, assessment, and equity arguments underlying standards-based reforms. An immense body of research from the 1980s and 1990s, later codified in *How People Learn* (National Research Council, 1999a), included new insights

about the nature of expertise, knowledge structures and connections to prior knowledge, the importance of metacognitive strategies and self-regulation, links between motivation and sense of self to what is learned, and even the emerging idea (then) that participation in social practices is a fundamental aspect of learning. Examples from this research base are myriad. Studies of learning in out-of-school settings, such as Collins, Brown, and Newman's (1989) cognitive apprenticeship, demonstrated the importance of situating abstract tasks in authentic contexts, very different from the inert and decontextualized forms of knowing required in schools. Lampert (1990) sought to shift classroom participation structures to more closely resemble standards of logical argument in the mathematical community. Cobb, Wood, and Yackel (1993) argued that new norms would need to be negotiated to overcome previously constructed norms about trying to guess the teacher's solution and avoiding evaluation. Luis Moll's "funds of knowledge" for teaching offered a practice whereby students' prior knowledge about agriculture, mining, economics, household management, medicine, and religion could be used as cognitive resources to engage students in more challenging and meaningful tasks (Moll, Amanti, Neff, & Gonzales, 1992).

Although the above ideas were widely shared among researchers, the NAEd report tried to explain to a popular audience what findings from cognitive and constructivist psychology meant for changes in classroom practice. For example, it was a mistake, left over from behaviorism, to postpone thinking and reasoning until after basic skills were learned by rote. The NAEd report also discredited widely held nativist beliefs about inherited abilities that lurked behind contemporary instructional practices, making it acceptable to reserve rich and engaging curricula for only an elite subgroup of students. Thus, the NAEd report endorsed, in principle, the idea of "high standards for all students" but noted that the needed changes to the system were unprecedented and monumental. The report cautioned further that despite hopes for greater opportunity and equity, reforms could actually exacerbate inequities if standards were accompanied by high-stakes assessments, if teachers in urban and poor school systems had the least access to professional development, and if students were punished for failures that were the system's fault. The authors tried to explain the apparent contradiction of knowing a great deal about learning and teaching but not having sufficient knowledge about how larger systems and social contexts could be sufficiently transformed to make the envisioned changes possible. The report emphasized the importance of capacity building, especially teacher professional development, and the need for continuous research and evaluation of the reforms' effects.

How naïve it was to imagine that policymakers' past practices would not trump ephemeral visions of reform. In 1994, two different versions of

standards-based reforms were installed in federal policy, Clinton's Goals 2000: Educate America Act and the reauthorization of the Elementary and Secondary Education Act (ESEA) called the Improving America's Schools Act (IASA). Both alluded to the systemic changes that would be needed, but it would be up to states to figure out how to make and fund those changes. IASA included principal and teacher professional development as part of Title II; but it was IASA's accountability mandates that determined its character and impacts. Subject-matter experts kept talking about research on learning, but when policymakers adopted rewards and sanctions as the drivers of change, standards-based reforms became an incentives theory of change. Mathematics education reformers developed beautiful examples of curricular resources that would help teachers help students develop deeper conceptual understandings; and learning-focused projects such as TERC and LRDC's Institute for Learning developed resources that could support transformative change. But none of these could have the reach of accountability mandates, which by definition affected every classroom and school.

By 1999 a National Research Council (NRC) report titled *Testing, Teaching, and Learning* (National Research Council, 1999b) concluded that the theory of action guiding standards-based reform might be overly optimistic because it assumed that teachers know how to educate all children to high levels of performance or know how to seek the relevant new knowledge. Accountability structures were thought to be sufficient to bring these extant resources to bear. Studies of what was happening on the ground, however, found that external accountability mandates landed in very different ways in rich and poor schools. Better-situated schools, as termed by Carnoy, Elmore, and Siskin (2003), that is, those serving more advantaged communities, were more able to respond coherently to accountability pressures. Better-positioned schools with relatively high "internal accountability" recognize that increased coherence around instructional practice required new curriculum content and new knowledge and skills for teachers and administrators—and found ways to change the structure of the work and gain those skills. Without this wherewithal, the reforms were a hallow shell.

The assessment strand of standards-based reforms has a similar history of grand hopes followed by erosion and misdirection. Evidence gathered in the late 1980s showed the negative effects of teaching to low-level tests. In particular, an important synthesis project led by Michael Feuer for the Congressional Office of Technology Assessment (U.S. Congress, 1992) brought together studies documenting the curriculum narrowing effects of high-stakes testing and resulting test score inflation, that is, test scores could go up without there being a corresponding increase in learning. In the United States, performance assessments were offered by

standards-based reformers as the remedy to these distorting effects. New assessments that better represented nobler learning targets would be "tests worth teaching to." In England and other countries in the United Kingdom, an Assessment Reform Group focused instead on formative assessments in classrooms as a potential counterforce to the damaging effects of school league tables. Building specifically on the important contributions of *How People Learn*, an NRC committee was formed to bring together new knowledge from research on both learning and measurement. The resulting NRC report, *Knowing What Students Know* (National Research Council, 2001), explained the different purposes of large-scale versus classroom-level assessments and how the two could be coherently linked by a shared model of learning.

From this foundational knowledge, the 1990s saw a brief flourishing of more inventive forms of assessment. These included portfolio assessments in Kentucky and Vermont and performance assessments in California and Maryland. But this heyday was short lived. Perhaps the most visible example was the California Learning Assessment System (CLAS) that lasted only 3 years. Religious groups objected to the content of reading passages and to the idea of the Sacramento bureaucrats assessing "critical thinking;" and newly elected policymakers resented the tradeoff that had been made, sacrificing individual student scores to make performance assessments possible (Kirst & Mazzeo, 1996). The real death knell to performance assessment reforms, however, came with the passage of the No Child Left Behind Act (NCLB) in 2001. The sheer volume of tests required and the mandate for individual student scores closed down any state testing program that had relied on matrix sampling to obtain school scores and made scoring of open-ended assessments cost prohibitive. NCLB required that every child be tested every year in reading and math from grades 3 to 8. Moreover schools would essentially be placed in receivership if they failed to demonstrate adequate yearly progress defined by increasingly out-of-reach targets. The idea that 100 percent of students would be expected to meet ambitious learning targets by 2014 was absurd on the face of it. By 2011, states began receiving waivers from some of the more draconian requirements, but this did not prevent a frantic, decade-long focus on raising test scores. NCLB also had an explicit focus on closing gaps between majority and minority groups, but its provisions failed to attend to the kinds of genuine learning opportunities that would make these leaps possible.

NCLB's relentless accountability pressures had further pernicious effects because of what the focus on test scores did to undermine the fledgling efforts being made to introduce formative assessment practices in the United States. In my address to the American Educational Research Association in 2000 titled "The Role of Assessment in a Learning Culture,"

I took up the formative assessment arguments of colleagues in the United Kingdom, Australia, and New Zealand and tried to draw connections between learning processes described from a Vygotskian perspective and what subject-matter experts were saying about ambitious learning goals. Numerous learning principles rendered from a cognitive perspective in *How People Learn* —attending to prior knowledge, substantive feedback, internalizing criteria, metacognitive benefits of self-assessments, teaching, and assessing for transfer—can also be taken up in socially supporting learning environments in ways that enable collaborative relationships between student and teacher (Gipps, 1999). In such a culture, developing an identity of mastery occurs as learners participate in a community of practice (Lave & Wenger, 1991). But these ideas cannot flourish in a test-driven environment.

In the wake of NCLB, entrepreneurs and test publishers co-opted the term formative assessment and used it to sell products to school districts with item formats that were cheap imitations of state tests. Another reform was undone by superficial understandings. Dylan Wiliam (personal communication, 2005) called these products "early warning summative tests." In essence, districts were paying good money for instruments that looked for all the world like teaching-the-test training materials. Patricia Burch (2006), who studies various forms of educational privatization, found that top testing vendors doubled their annual sales between 2000 and 2006. A group of scholars brought together under the auspices of the Council of Chief State School Officers issued a formal decree explaining why more frequent testing with mostly multiple choice items bore no resemblance to the learning research supporting formative assessment. The very tiniest victory was won when the term interim (Perie, Marion, & Gong, 2009) or benchmark assessments was adopted instead of formative assessments to describe formal tests given every 4 to 6 weeks. However, the use of interim assessments themselves in no way abated.

Simply examining the most popular of these commercial test products should have made it clear why they are so unlikely to produce deep and meaningful changes in learning opportunities. For the most part, though computer delivered, they look just like the narrow basic skills tests from the 1980s. They were not developed to provide diagnostic insights about students' understandings. Empirical studies examining the use of such measures show that instructional responses are largely procedural or at best they only prompt teachers to try something different (Nabors Olah, Lawrence, & Riggan, 2010). The few positive examples of benchmark assessment results being more deeply linked to instruction improvements appear to be led by highly committed principals or teacher leaders, but the more prevalent practice is item-by-item reteaching with little attention to student thinking (Blanc, Christman, Liu, Mitchell, Travers, & Bulkley, 2010; Shepard, 2010; Shepard, Davidson, & Bowman, 2011). In our study of interim assessments in seven districts in two states, we found that teachers described not a learning culture but a benchmark assessment or accountability culture exemplified by posting students' scores in the hallway and giving feedback to students in terms of how many more items they needed to score correctly to reach proficiency (Shepard et al., 2011).

The NAEd white paper on *Standards, Assessments, and Accountability* (Shepard, Hannaway, & Baker, 2009) included some of this same history on standards-based reforms as well as a summary of policy research documenting limited investments in capacity building. Effective examples of teacher professional development were cited where they did occur. The NAEd working group authors made several recommendations about needed changes within the existing federal accountability framework, calling for well-articulated learning progressions, ambitious but realistic learning targets, ongoing evaluation of accountability systems to ensure that they contribute to school improvement, and closer investigation of school performance before imposing remedies or sanctions. In my view, however, the most important of our recommendations was the following: "The federal government should support an intensive program of research and development to create the next generation of performance assessments explicitly linked to well-designed content standards and curricula."

In this one recommendation are two critically important ideas: first, that an intensive program of research is needed to develop next-generation assessments and second, that performance assessments and curricula should be developed together.

Given the theme of this essay—that reforms are undone by superficial understandings or by hollow enactments of idealized schemes—it should not be surprising that the idea of an intensive assessment research and development (R&D) effort was undermined, essentially by the decision to deliver new operational tests on a broad scale in too short a time. The Department of Education heard the argument that state consortia would be needed to build and try out the kind of curriculum-linked, learning-progression-linked assessments outlined in the NAEd white paper and in *Knowing What Students Know* a decade before. They understood that individual states would not be able to undertake such challenging development on their own. However, the distinction we had drawn between "the political process needed to achieve consensus and guide policy decisions versus the scientific expertise needed to develop and rigorously evaluate curricular materials, instructional strategies, and assessments" (Shepard, Hannaway, & Baker, 2009, p. 8) was lost.

Following the Great Recession, the American Recovery and Reinvestment Act monies made it possible for the federal government to invest in

developing next-generation assessments. Expert testimony was sought, which forewarned of all the past problems, but sometimes promised grandly how these problems would be overcome by the affordances of technology. The resulting Race to the Top call to develop Comprehensive Assessment Systems (U.S. Department of Education, 2010) was layered with enormous demands requiring that consortia comprising at least 15 states develop measures of college- and career-ready, cross-grade achievement trajectories in partnership with higher education institutions. Proposers were expected to correct all of the shortcomings of past assessments: measure the full range of performance implied by the standards, including aspects of achievement that had heretofore been difficult to measure; elicit complex applications of knowledge and skills; measure accurately for high- and low-achieving students; and so forth. The successful consortia were also required to ensure that their assessments were "valid, reliable, and fair." Professionals involved in the consortia performed herculean tasks, but there was no way to live up to all of the rhetorical claims, and some shortcuts were necessary. For example, validity analyses had to rely more on content reviews by experts and internal statistical properties of the assessments during field trials rather than empirical studies of assessment results linked to either school improvement or student success in college and career.

The two consortium tests, Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced, were rolled out in spring of 2015. It is fair to say that these new assessments are generally of higher quality than past state assessments because they are targeted at higher levels of thinking and use more open-ended formats. They are not so good, however, as what might have been possible if investments had been made in a genuine R&D effort. It is also fair to say that both consortia have made some significant missteps, assuming too much about computer availability in all schools, requiring too much testing time, and sometimes using drag and drop and other technological interfaces in ways that hurt rather than enhance representations of important content. Consortium leaders are still trying to solve too many irreconcilable problems. Broad coverage with open-ended performance assessments would be possible with reasonable amounts of testing time if policymakers would reconsider the possibility of matrix sampling and roll back the demand for test-based teacher accountability. More importantly, policymakers might also recall, from the long history of standards-based reforms, that top-down mandates do not help poor schools get better if educators in these schools do not have access to resources to teach in fundamentally different ways.

In the fall of 2015, just before this essay is to be published, results will be released showing miserable student performance on PARCC and Smarter Balanced. No amount of explaining will help the public understand the ambitions of the new content and practices frameworks, the stringency of proficiency cutoffs, nor the lack of resources or time to turn an entire system toward these new ends. Public schools will be bashed again and dedicated, caring teachers will continue to leave the profession in droves.

I began this account by asking why educational reforms are not successful, given that we know so much from research on learning. The answer and lesson to be learned by researchers as well as policymakers is that *cheap, superficial, and coercive versions of reform ideals will inevitably prevent deeply substantive, hoped-for changes*. The kinds of transformative changes that are needed—to make real differences in learning opportunities—are difficult and cannot be made on command. No amount of talk about "capacity building" can substitute for the supports that are needed. As predicted, inequities are increased when short-cut strategies are the best that can be done in response to accountability pressures. Researchers who helped conceptualize the beginning of the standards movement in the 1990s wanted to create policy coherence at the top that would support meaningful changes at the bottom of the system. But oppressive regimes at the top only create scurrying at the ground level. That is why drilling—on worksheets or interim measures that imitate accountability tests—has been so much more pervasive than deeper changes in curriculum or instructional practices.

We are now in a world of next-generation, Common Core State Standards (or new state standards that closely resemble CCSS), and goals such as critical thinking and problem solving now have wider appeal. If policy leaders want to support more profound changes in teaching and learning opportunities—in poor as well as rich schools—then they will need to reconsider the juggernaut of accountability testing. To do this it might be helpful to return to the recommendations from *Knowing What Students Know* and recall the very different purposes of large-scale assessments for monitoring and accountability versus classroom-level tests to inform teaching and learning. There will surely need to be refinements of PARCC, Smarter Balanced, and various other state tests. Ideally they would be used to collect data and track progress, not to create incentives by imposing unreasonable targets. If leaders insist on targets, then they should be informed by what Bob Linn (2003) called an "existence proof," that is, high standards that at least someone has been able to reach; for example, schools might be asked to raise achievement to the levels currently attained by the top 25 or 10 percent of similarly situated schools.

Not to be forgotten, an important and distinctly different need is for the development of curriculum materials to support teachers in learning to teach to much more ambitious standards. The design of assessment tasks, both formative and summative, should be an integral part of

such curriculum development. The National Science Foundation's current funding of learning progressions in science is one way to study jointly the furthering of learning at the same time that we get better at assessing and interpreting student thinking. If we have learned only one thing from the disappointments of standards based reforms, then it should be that trying to install giant systems is a mistake. It would be much better to take a step back from the most aversive aspects of current accountability systems and focus instead on smaller scale projects with adequate time to learn from mistakes and improve. Then we could imagine implementing such curricular materials and next-generation assessments on a larger scale, but only if at each step we have evidence that systems are becoming more equitable, not less.

References

Blanc, S., Christman, J. B., Liu, R., Mitchell, C., Travers, E., & Bulkley, K. E. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education, 85*(2), 204–225.

Burch, P. (2006). The new educational privatization: Educational contracting in the era of high stakes accountability. *Teachers College Record, 88*(2), 129–135.

Carnoy, M., Elmore, R., & Siskin, L. S. (2003). *The new accountability: High schools and high-stakes testing*. New York: Routledge Falmer.

Cobb, P., Wood, T., & Yackel, E. (1993). Discourse, mathematical thinking, and classroom practice. In E. A. Forman, N. Minick, & C. A. Stone (Eds.), *Contexts for learning: Sociocultural dynamics in children's development* (pp. 91–119). New York: Oxford University Press.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453–494). Hillsdale, NJ: Erlbaum.

Gipps, C. V. (1999). Socio-cultural aspects of assessment. In P. D. Pearson & A. Iran-Nejad (Eds.), *Review of Research in Education* (Vol. 24, pp. 355–392). Washington, DC: American Educational Research Association.

Goals 2000: Educate America Act. 20 U.S.C. § 5801, et seq. (1994).

Improving America's Schools Act, 20 U.S.C. § 6301, et seq. (1994).

Kirst, M. W., & Mazzeo, C. (1996, April). *The rise, fall and rise of state assessment in California, 1993–1996*, paper presented at the annual meeting of the American Educational Research Association, New York.

Lampert, M. (1990). When the problem is not the question and the solution is not the answer: Mathematical knowing and teaching, *American Educational Research Journal, 27*(1), 29–63.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.

Linn, R. L. (2003). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives, 11*(31). Retrieved from http://epaa.asu.edu/epaa/v11n31.

McLaughlin, M. W., & Shepard, L. A. (1995). *Improving education through standards-based reform. A report of the National Academy of Education Panel on Standards-Based Reform*. Stanford, CA: National Academy of Education.

Moll, L. C., Amanti, C., Neff, D., & Gonzalez, N. (1992). Funds of knowledge for teaching: Using a qualitative approach to connect homes and classrooms, *Theory Into Practice, 31*(2), 132–141.

Nabors Olah, L., Lawrence, N. R., & Riggan, M. (2010). Learning to learn from benchmark assessment data: How teachers analyze results, *Peabody Journal of Education, 85*(2), 226–245.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

National Research Council. (1989). *Everybody counts: A report to the nation on the future of mathematics education*. Washington, DC: National Academy Press.

National Research Council. (1999a). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

National Research Council. (1999b). *Testing, teaching, and learning: A guide for states and school districts*. Washington, DC: National Academy Press.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28*(3), 5–13.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for education reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37–75). Boston, MA: Kluwer Academic.

Shepard, L. A. (2010). What the marketplace has brought us: Item-by-item teaching with little instructional insight. *Peabody Journal of Education, 85*(2), 246–257.

Shepard, L., Hannaway, J., & Baker, E. (Eds). (2009). *Standards, assessments, and accountability*. Washington, DC: National Academy of Education.

Shepard, L. A., Davidson, K. L., & Bowman, R. (2011). *How middle-school mathematics teachers use interim and benchmark assessment data*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Smith, M. S., & O'Day, J. (1990). Systemic school reform. In S. H. Fuhrman & B. Mahen (Eds.), *The politics of curriculum and testing: The 1990 yearbook of the Politics of Education Association* (pp. 233–267). London, UK: Taylor & Francis.

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

U.S. Department of Education. (2010). Overview information: Race to the Top Fund Assessment Program; Notice inviting applications for new awards for fiscal year (FY) 2010. *Federal Register, 75*(68), 18171–18185.