C-R-E-S-S-T

National Center for Research on Evaluation, Standards, and Student Testing

Technical Report

You can view this document on your screen or print a copy.

- UCLA Center for the Study of Evaluation

in collaboration with:

- University of Colorado
- NORC, University of Chicago
- LRDC, University of Pittsburgh
- The RAND Corporation

Old and New Beliefs About Measurement-Driven Reform: "The More Things Change, the More They Stay the Same"

CSE Technical Report 373

Audrey J. Noble and Mary Lee Smith
CRESST/Arizona State University

April 1994

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
Graduate School of Education
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

OLD AND NEW BELIEFS ABOUT MEASUREMENT-DRIVEN REFORM: "THE MORE THINGS CHANGE, THE MORE THEY STAY THE SAME"[1]

Audrey J. Noble and Mary Lee Smith

CRESST/Arizona State University

This article reports on a study of a new performance-based test mandate in the state of Arizona. The measurement-driven reform provided an opportunity to examine the interplay of policy and practice, specifically the use of testing to effect educational change. Attempts to reform schools through mandating assessments are categorized as behaviorist and cognitive-constructivist. Beliefs and assumptions underlying each model are analyzed. We argue that recent attempts to meld cognitive-constructivist beliefs about pupil learning with behaviorist views of reforming schools and teaching teachers are contradictory.

The current enthusiasm for reforming schools has no shortage of players. Local organizations and professional groups (Coalition for Essential Schools, National Council of Teachers of Mathematics, the Network of Accelerated Schools, the Center for Establishing Dialogue in Teaching and Learning) attempt to transform schooling from the bottom up or from the inside out. Yet these local efforts are overshadowed by state and federal projects to reform schools from the top down, for example, by lengthening the school day and year, setting more rigorous and extensive graduation requirements, fully funding Head Start, and introducing market incentives for schools. The most pervasive tool of top-down policy reform is to mandate assessments that can serve as both guideposts and accountability mechanisms. Various state and federal projects seek to drive reform by means of high-profile, high-stakes tests. Their backers envision not simply improved performance on traditional goals but alternative conceptions of curriculum and pedagogy.

[1] **Thanks to Gail Hackett, James Powell, and Gene V. Glass for reacting to early drafts and to our co-researchers and research participants for their contributions.**

In this paper we explore assumptions and issues of psychological, pedagogical, and reform theories that undergird measurement-driven reform (MDR). We draw on our previous (Smith, Edelsky, Draper, Rottenberg, & Cherland, 1990; Smith, 1991) and current research (Noble, 1993; Smith, Noble, Heinecke, Cabay, Junker, & Saffron, 1993) on mandated assessment in Arizona as well as extant theoretical and empirical work and the rhetoric of reformers.

Old View of Measurement-Driven Reform (MDR)

Popham outlined the traditional notion of measurement-driven instruction (MDI): Assessments direct teachers' attention to the content of test items, acting as powerful "curricular magnets" (Popham, 1987). In high-stakes environments, in which the results of mandated tests trigger rewards, sanctions, or public scrutiny and loss of professional status, teachers will be motivated to pursue the objectives the test embodies.

The traditional view is grounded in a behaviorist psychology and pedagogy. The desired performance of pupils is brought about by reinforcing successive approximations of correct performance. Academic tasks are broken down into discrete units and presented to the pupils. Rewards and progress through an ordered hierarchy of tasks and skills are contingent on correct performance. Inadequate responses result in repetition and repeating through the same material until it is mastered, with instruction in "higher order" skills resting on a foundation of "basic skills." In the behaviorist model, the pupil is considered a passive recipient of knowledge. The intentions of learners are generally ignored. Furthermore, teaching teachers follows the same model of pedagogy as teaching pupils. Whatever the system's goals and objectives becomes the content that teachers deliver.

Beliefs about testing follow from beliefs about teaching and learning. Whatever the objectives cover, tests can be written to measure them, can be administered to all, and re-administered as necessary to those who fail. The primary values attached to tests include content validity (whether they cover the objectives), objectivity (will they be scored alike by two readers), standardization (are they the same for all students), and comparability. In practice, assessment that is consistent with these beliefs has most often been multiple-choice in form; that is, students respond to common questions by recognizing answer options

already determined to be correct. Teaching to the test is simply considered good instruction if test items mirror the given objectives.

In this traditional view, beliefs about reform of schools are merely an extension of behaviorist psychology: Teachers and other professionals respond according to the contingencies placed on them in the form of rewards for following mandated standards and achieving high scores on mandated assessments. They act rationally and acquiesce to the reforms so as to avoid sanctions and punishments or loss of status that would follow publication of low test scores. As a consequence, schools should improve by the introduction of high standards supported by mandated assessments. The behaviorist view of reform is what Combs (1991) referred to as "closed system," in which the goals or ends are mandated, the means to the ends are assumed or ignored, participants are passive, and conformity is valued. Take as an example the hypothetical state goal "Students will understand inferential statistics." The behaviorist MDR assumes that the statement of the goal and its measurement in a mandated test provide the target for which schools should strive and symbolize the social value associated with its accomplishment. This is the "end" in the closed system model. The "means" to the end, that is, providing appropriate texts and materials, making room in the official and enacted curriculum, hiring teachers who themselves understand statistics or retraining those who currently do not, and, most importantly, having the will and intention to change, are overlooked or unspecified.

Are the behaviorist beliefs underlying measurement-driven reform warranted? A small body of evidence addresses the functions of assessments from the traditional viewpoint. Popham's (1987; Popham, Cruse, Rankin, Sandifer, & Williams, 1985) own studies showed that performance improves on second and third administrations of high-stakes tests, suggesting positive effects. A study by Koretz, Linn, Dunbar, and Shepard (1991), however, concluded that the positive effects on the specific achievement test employed in a high-stakes environment may not generalize to alternative tests and indicators of achievement. Thus, according to Linn (1993, p. 3), "considerable caution is needed in using achievement test results to draw inferences about the quality of education." In light of this, belief in the power of assessment to reform schools must be tempered.

Moreover, research on the impact of mandated testing shows effects contrary to the intentions and expectations of those who would reform schools. High-stakes testing shapes the curriculum, but not necessarily in straightforward ways. Corbett and Wilson (1991), in their examination of statewide testing programs in Pennsylvania and Maryland, found that the higher the stakes, the more likely narrowing of the curriculum will occur. Smith et al. (1990; Smith & Rottenberg, 1991) found that schools neglected topics such as science, social studies, and writing that the mandated tests failed to cover. In the schools Mathison (1987) studied, the test became the curriculum. Darling-Hammond and Wise (1985) found that teachers emphasized the exact contents of the mandated test rather than the underlying concepts and goals around which the test was constructed. Herman and Golan (1991) found that teachers adjusted the sequence of their curriculum based on what was included on the test.

High-stakes mandated testing also has been shown to influence pedagogy. Mathison (1987) discovered that teachers altered their instructional materials to resemble the format of mandated tests. Stodolsky (1988) found that the existence of mandated testing led teachers to neglect team teaching approaches. Of the teachers Shepard and Dougherty (1991) surveyed, 69% reported giving more emphasis to teaching basic skills because of the content of mandated standardized tests; half of the respondents admitted to slighting subjects such as science, social studies, and higher order thinking skills that the tests did not cover. Romberg, Zarinnia, and Williams (1989) found that, as a result of the emphasis on test results, teachers increased attention to paper-and-pencil computation and decreased attention to project work, activities that involved use of calculators and computers, and cooperative learning activities.

In another study (Smith et al., 1990), teaching became increasingly test-like. Teachers used worksheets cast in multiple-choice form, spelling was redefined as practice on identifying which of three words was misspelled, and "problem solving" was taught and tested as following a five-step algorithm for converting story problems to computations. All of these methods matched the mandated test format. McNeil (1986) concluded that mandated competency testing engendered low-quality, mechanistic instruction.

Research shows that mandated tests also put a heavy burden on instructional time. In one district Smith et al. (1990) found the battery of

mandated tests, plus time taken in test preparation and recuperation from testing, decreased instructional time by about 20 days in one year. A survey of Arizona schools (Nolen, Haladyna, & Haas, 1989) showed that nearly one-third of the elementary school teachers started to prepare their pupils two or more months before the test schedule. All such activities replace teaching with testing. Test preparation is often intense and its effects problematic (Haladyna, Nolen, & Haas, 1991; Mehrens & Kaminski, 1989; Shepard & Dougherty, 1991; Smith, 1991). Studies by Cannell (1987) and Linn, Graue, and Sanders (1990) showed that standardized achievement test scores, used in high-stakes conditions, increased after initial administration, eventually contributing to the Lake Wobegon effect, wherein the test score averages of most states exceeded the national average. Authors have attributed this apparent anomaly to old norms or perhaps inappropriate test preparation.

High-stakes testing affects teachers directly and negatively. For example, Smith and Rottenberg (1991) found that the emphasis on test results diminished teachers' sense of themselves as autonomous professionals and authorities on instruction and curriculum. The dictates of externally mandated tests reduced both their perceived levels of professional knowledge and status (Shepard & Dougherty, 1991). A study by Hatch and Freeman (1988) revealed that teachers reported considerable distress because of the conflict between instructional methods they felt forced to adopt and their own beliefs about children's learning needs. Consequently, classroom instruction defined by high-stakes tests rather than by teachers had the effect of driving out good teachers and "de-skilling" those who remained. Good teachers either found a means to resist the de-skilling process or left teaching (McNeil, 1986). Fish (1988) concurred that the more pressure teachers felt to raise test scores, the lower their professional self-images.

Research also suggests that pupils suffer in the pressurized environment of high-stakes testing. Teachers report that young children experience anxiety, frustration, physical symptoms, loss of self-esteem, and give up without trying on the test (Nolan et al., 1989; Smith, 1991). Minorities and disadvantaged pupils are penalized when test scores affect placement and teaching decisions (e.g., retention, special education placement, tracking) (Holmes, 1989; Medina & Neill, 1988; Shepard, 1989, 1991). McGill-Franzen and Allington (1993) found that schools in high-stakes environments used retention and special education

placement as ways to take lower-scoring pupils out of the test score pool so that the schools' average scores would look better. Madaus, West, Harmon, Lomax, and Viator (1992) found that the higher the percentage of minority students, the more teachers valued test scores and oriented to them. Test results influenced decisions about student progress, student placement in special programs, curriculum planning, and textbook selection. Teachers in classrooms with a high number of minority students, feeling pressure to improve scores, focused on low-level skills: test-taking skills and tested topics.

High-stakes testing has effects that may run counter to the true intents of those who wish to improve schools. Old MDR fails to take into account the intentions of educators and how statewide policies often are reinterpreted at the local level (Lipsky, 1980). Narrowing of curriculum, degrading of teaching method, and intensive test preparation occur because educators intend to maximize gains on highly fallible and susceptible indicators (Smith, 1991). What can be measured is usually just a small sample of the domain of interest, not necessarily representative of the goal. But high stakes are attached to the test score rather than the goal-directed activity. Suppose that the goal is understanding statistics and the measure covers the *t*-test. The available research evidence suggests that schools will focus *only* on the *t*-test and neglect not only alternative, untested areas of mathematics but also closely related but untested content like analysis of variance. Moreover, teaching will mimic the form of the test, with most attention paid to routine computation of *t*-test data and recognition of correct alternatives and with little attention paid to application or problem solving. Teaching test-taking skills and drilling on multiple-choice worksheets is likely to boost the scores but unlikely to promote general understanding. Using test scores to monitor the progress of schools equates a narrow indicator with a comprehensive view of what is being accomplished. As a result, by focusing on improving test scores, only test scores, and not schools themselves, will improve.

New View of Measurement-Driven Reform (MDR)

Faced with these complications, policy makers and scholars who still believe in the power of assessment to drive reform and change schools have focused on the fallacies in the psychology and pedagogy of the traditional view as well as the form of the measurement itself. Resnick and Resnick (1989a) asserted three

principles of accountability assessment: (a) You get what you assess; (b) you do not get what you do not assess; and (c) You should build assessments toward how you want educators to teach. From the first and second principles, which can be inferred from research, they reached the third: that high-stakes assessment could drive reform if it followed better psychology and pedagogy and employed more appropriate measurement forms, namely performance-based assessments. If tests affect curriculum and instruction, the argument goes, performance-based assessment could serve as an impetus for a thinking-oriented curriculum geared toward developing higher order abilities and problem-solving skills (Honig, 1987; Resnick & Resnick, 1989b). Instruction directed toward preparation for a performance-based assessment promotes better instructional practice (Baker, Aschbacher, Niemi, & Sato, 1992). A better test will produce better results. Teaching to the test, accepted by scholars as inevitable and by teachers as necessary, becomes a virtue, according to this line of thinking.

The psychology and pedagogy of the new view, which we oversimplify by labeling it cognitive-constructivism, rejects behaviorism as a model of learning. Instead, it emphasizes three interrelated dimensions: (a) Learning is a process of construction; (b) it is knowledge-dependent; and (c) learning is situated in a context (Resnick, 1989a).

Viewed as a process of construction, learning is not an act of recording discrete bits of information, each bit independent of the others and needing only to be repeated until it is mastered. Instead, it is a process of interpretation and construction of meaning (Glaser & Bassok, 1989). Students are active participants in their learning, constructors of knowledge, not passive recipients of information and skills (Piaget, 1948/1974). For learning to occur, students need the opportunity to use information and experience its effects upon them (Bransford & Vye, 1989). Moreover, learning is intentional. The meanings of the teaching and learning act must be taken into account. One cannot assume that an instructional activity means the same thing to one pupil or teacher that it does to another. Variation in meaning and intention are significant and need be acknowledged by anyone interested in measuring the effects of that interaction.

The second key consideration of the cognitive-constructivist perspective is that learning is knowledge-dependent. It is not merely an act of receiving information but one of interpreting information through earlier learning. The

role of prior knowledge and experience is given special attention, for what a pupil learns on a given occasion is dependent on what has been already learned. It is much easier for individuals to learn more about their areas of expertise than it is for them to learn about other topics outside their experience. Chi (1978) compared 10-year-old children and adults on their ability to remember chess positions and digit tasks. As was expected, the adults performed better than the children on the digit memory task. But in the chess task, the children's memory for the placement of the pieces was far superior. What made the difference was that the children in the group had played tournament chess while the adults had little prior experience with the game. The cognitive view also asserts that students rarely learn usable knowledge as a result of simply being told. "People continually try to understand and think about the new in terms of what they already know" (Glaser 1984, p. 100). For knowledge to become generative, it must be used again and again as a means of interpreting and explaining new information (Resnick, 1989b).

The third dimension of the cognitive-constructivist view reveals that knowledge is not independent of the context in which it develops. Much of the earlier thinking about context-independent learning is based on the belief that if something is learned in one context it can be readily translated into another. Specialized programs to teach students study skills, problem solving, and reasoning represent efforts that define these skills as general rather than context-dependent (Resnick, 1987). However, the constructivist perspective questions whether learning can be separated from the context in which it occurs (Gergen, 1985; Resnick, 1987). Learning not only occurs in a context but is also social. Group learning experiences provide children social support and encouragement along with exposure to alternative views. Each student contributes and benefits from the participation of others (Glaser & Bassok, 1989). For example, in the study of literature, students may participate in dialogue groups. Although each child reads the same book, the focus is not on coming up with the same answers at the end of the chapter. Students not only share their own thoughts and impressions, stemming from their experience with similar information; they also create new understanding as they interact with other students. The focus is on the process, which develops over time, as much as the end product.

The cognitive-constructivist perspective proposes that effective instruction must mesh with how students think (Dewey, 1963; National Council of Teachers of Mathematics, 1989). The direct instruction model, the "tell-show-do approach," does not match how students learn (Baroody & Ginsburg, 1990), nor does it take into account pupil intention, interest, and choice. Teachers must put away their familiar and standard methods: basal readers, worksheets that drill on isolated basic skills, mastery learning, and grade-level- and subject-matter-specific texts and materials. Teaching that fits this cognitive-constructivist view of learning is likely to be holistic, integrated, project-oriented, long-term, discovery-based, and social. The thinking curriculum emphasizes learning for understanding. Teachers need to recognize and comprehend that students bring their own knowledge to each learning encounter. The challenge is to recognize that learning a new concept may involve confronting old, and sometimes contradictory, beliefs. Students need assistance in constructing new theories (Resnick, 1989a). If they are unable to integrate or replace the old with the new, they are likely to forget the new information and rely on their older perceptions (Driver, Guesne, & Tiberghien, 1985). For example, if students first learn reading phonetically, they learn to define it as a process of linking sounds to symbols. However, when confronted with the concept of reading for meaning, students may struggle with their earlier beliefs. When they encounter an unfamiliar word, they may revert to sounding it out. Unable to integrate the new concept of reading for meaning in a context, they fall back on their old perceptions and phonetic strategies and are no closer to understanding the word.

The cognitive-constructivist view of psychology and pedagogy aligns with a mode of assessment known as *performance* or *alternative*. It rejects as inappropriate the sole use of traditional multiple-choice items that test isolated bits of knowledge and skills. This type of assessment has been described as decontextualized and decomposed (Resnick & Resnick, 1989a). For example, a traditional mathematics problem asks the student to divide one amount by another.

$26.94 / 3 =

A. $13.47 **C. $8.83**

B. $8.98 **D. none of the above**

In contrast, performance tests allow pupils to construct responses to questions, topics, and problems on more realistic and complex texts. The alternative presentation of the same problem might appear as follows:

> **Three fourth-grade teachers at Park City Elementary School decided to take all their students on a picnic. Mr. Clark spent $26.94 for refreshments. Since the three teachers wanted to share the cost of the picnic, Mr. Clark used his calculator to determine that each teacher should pay him $13.47. Is his answer reasonable? Explain.**
>
> **(National Council of Teachers of Mathematics, 1989)**

Thus, cognitive-constructivists see performance tests as a form of testing that parallels their view of how pupils learn and should be taught.

To refashion this particular form of assessment to serve both accountability and instructional improvement functions is a psychometric as well as a political undertaking. The key is to add the values of comparability and objectivity to the responses of pupils. That is, the test activities and the scoring procedures must be comparable and standardized, while preserving the holistic properties of good performance tests. Tests must be mandated by government agency. The aggregated results must be highly visible and of consequence. Those who support the constructivist view of measurement-driven reform argue that, since you get what you assess, by using performance tests in the *function* of high-stakes accountability, the thinking curriculum and the challenging state and national standards can be achieved as proposed.

Are the Beliefs Underlying New Measurement-Driven Reform Warranted?

Few empirical studies exist of the use and effects of performance testing in high-stakes environments. Koretz, Stecher, and Deibert (1992) found in their study of the Vermont portfolio assessment project that scoring reliability was sufficient to support inferences about achievement at the state level but not at the school or district level. From teachers' reports, however, they concluded that the testing program had effects on curriculum and pedagogy that were consistent with the state's efforts to reform schools. Torrance (1993) described the United Kingdom's efforts to implement its National Curriculum through the use of a National Assessment. The assessment program (which, in many ways,

mirrored the Arizona experience, described in the next section) introduced "instructional packages" for teachers' use to push them toward the desired curriculum, pedagogy, and tests of achievement. The performance tests were administered and scored to provide comparable and public data. Torrance found that the complexity of the tasks, the emphasis on curriculum "delivery," the absence of appropriate professional development, the resource and time demands of the programs on teachers, the "psychometric imperatives" for standardization and comparability, the high-stakes nature of the assessment, and the limited time and budget to carry out the program produced effects possibly contrary to the intent of the reform.

The Arizona Case

Prior to 1990, mandated assessments in Arizona mirrored the traditional view of measurement-driven reform. By legislative act, schools tested every child, every year using standardized, norm-referenced tests (the Iowa Tests of Basic Skills, for example) as well as criterion-referenced assessments. Results were published by school and grade level, and newspapers ranked schools according to test results. The pressure of the high-stakes assessment led many districts to align their entire curriculum to the standardized tests and spend inordinate time in preparation for them (Haas, Haladyna, & Nolan, 1989). However, some constituencies were dissatisfied with the norm-referenced test, concerned that it only covered a fourth of the state's legislated curriculum framework (i.e., the Arizona Essential Skills) and promoted inappropriate test preparation. These stakeholders combined forces with those who opposed the test because of its deleterious effects on students, teachers, curriculum, and instructional time. Thus, the alliance resulted in legislative action to change the test mandate. The result was the Arizona Student Assessment Program (ASAP), which (a) reduced standardized testing to three grades, (b) encouraged more options for district testing, and (c) added performance assessment in Grades 3, 8, and 12. The fourth component of ASAP was a requirement for a District Assessment Plan. School districts were required to demonstrate how their curriculum aligned with the state Essential Skills, along with how they intended to track student mastery of those skills. ASAP as a package, therefore, addressed several sets of stakeholders. However, two constituencies emerged: those who wanted state testing to exert pressure toward more cognitive-

constructivist, integrated curriculum and pedagogy and those who wanted to promote school and district accountability.

The performance test component of ASAP is described as more holistic assessment in that reading, mathematics, and language skills are tested in an integrated form. The test is said to reflect constructivist learning theory because pupils construct answers to test questions rather than respond to multiple-choice options. Another claim is that the test assesses higher order problem-solving abilities rather than rote memory or simple response to items representing isolated skills. Teaching to this test is viewed as desirable rather than distorting. Because of these features, the performance test is said to be state-of-the-art and consistent with "the best we know about how children learn." The Arizona Department of Education (ADE) believes that the existence of ASAP would cause teachers to emphasize the state Essential Skills and would move teachers in the direction of a more cognitive-constructivist pedagogy and integrated curriculum.

Following a one-year of pilot administration and scoring and a single study of the reliability and validity of the performance test, the ADE implemented the ASAP program throughout the state. The short time frame and minimal budget caused confusion and frustration in that first year, particularly in schools and districts that were not already practicing cognitive-constructivist pedagogy. Limited state-funded professional development activities focused on training teachers to score the assessments. Some districts provided instruction for teachers to change their teaching; others did not, according to their resources or their inclinations.

Similar to the experience in the United Kingdom (Torrance, 1993), the ADE relied on sample assessment forms to change instruction in the desired direction. More than 200 different forms of reading, writing, and mathematics performance-based tests were distributed to teachers. Each test was designed to measure one or more Essential Skills. One third-grade assessment, "interpreting word problems," purported to measure a cluster of seven skills (e.g., interpreting word problems by using role playing, adding and subtracting two 3-digit whole numbers, using informally the properties of commutativity, associativity, and identity, and writing mathematical sentences to represent a situation). Teachers were encouraged to use the alternative test forms for curriculum alignment and classroom instruction. Districts were required to indicate student progress

toward mastery of the Essential Skills at every grade. For this phase of ASAP, districts had the option of reporting results on either an alternative form of the performance test, district criterion-referenced tests, or portfolio assessments.

The ADE attempted to combine two assessment functions, instructional improvement and accountability. To make the results more comparable, efforts toward standardization prevailed. The test went from being a "social" and "process" assessment to an individual and "point in time" test. All students in Grades 3, 8, and 12 were tested on the same days. All students, according to grade level, took the same test. Test administration was timed, teachers' role as mediator was restricted, and students working collaboratively became redefined as "cheating." To make scoring of the tests more reliable and less "subjective," ADE constructed scoring criteria. According to some teachers, this action resulted in effectively reducing the complexity and the richness of the responses into simplistic 5-point scales.

Results from the pilot study and the first year of statewide administration appeared in a format similar to standardized test scores. Summary of assessment results by Essential Skills group, such as "writes a report based on personal observation," illustrated a frequency distribution of student scores and number of students participating (N). Scores were reported according to mean, median, standard deviation, and range. The results were also given by gender, special program membership (i.e., special education, bilingual, Chapter 1), and race/ethnicity. Although the form of assessment was performance-based, its function was high-stakes. After the first year of statewide implementation, district scores for reading, writing, and mathematics appeared on the front pages of Arizona newspapers. Headlines included indictments such as "State's pupils losing numbers game," "Tests say schools are failing," and "Math scores in state distress officials." Performance assessment entered the arena of accountability and reform, consistent with the intentions of the "new" measurement-driven reform movement.

The Interplay of Policy and Practice

The purpose of our study (Noble, 1993, Smith et al., 1993) was to portray how the ASAP reform was conceived, negotiated, and implemented and to document initial responses to it. We focused on the beliefs that influenced the change in the testing mandate at the policy-making and administration levels

and those that influenced educators' reactions to and implementations of it. Interviews with key policy makers and stakeholders and analysis of documents provided evidence at the macro-level. A multiple case study, encompassing year-long classroom observation, informant interviews, and focus group interviews at four elementary schools provided information at the micro-level. From the latter, we developed an understanding of how different practitioners in different kinds of setting reacted to the testing mandate and what effects it had on curriculum and instructional practice. Our research approach was multi-perspectival, and our frameworks were taken from Rein (1976) and Lipsky (1980). We were particularly interested in examining the interplay between policy and practice. One major facet of our analysis followed Erickson's (1986) suggestion of deriving assertions through induction from the corpus of data that could then be tested by searching for confirming and disconfirming instances, conducting negative case analysis, writing vignettes, and the like. In such a large and complex study, many such assertions can be generated. Space limitations prevent us from presenting more than one in this report.

The following assertion organizes much of our data. The Arizona Student Assessment Program as an instance of measurement-driven reform is internally inconsistent. While ideally the program promotes cognitive-constructivist assumptions about psychology and teaching of pupils, logistically it is built on behaviorist assumptions about reforming schools and teaching teachers. The learning principles promoted for students were ignored as they applied to teachers. Moreover, the mandate, as a closed system of the reform, fostered a behaviorist view of change. As reform instruments, mandates such as ASAP specify targets or ends, ignore means, demand compliance and punish noncompliance, and thereby attempt to create uniformity of practice (McDonnell & Elmore, 1987). Interviews with ASAP policy makers confirm these expectations of compliance and uniformity. Past legislative efforts to align instruction with state curriculum frameworks had proved somewhat ineffective in bringing about compliance. One policy maker from the Arizona Department of Education spoke of schools' responsiveness to mandates:

> **I think in the state over the years perhaps there was a culture of not paying too much attention to mandates or legislation or that kind of thing. We came out with the Essential Skills. We said this was part of the new law and people interpreted that in keeping with the way they did things normally in their school districts—some of**

> them looked at it very seriously and began to change right way—others thought "this may go away."

With a focus on compliance, the legislature added the assessment dimension to the mandate:

> They [districts] weren't necessarily focusing on those basic skills or those Essential Skills. I think part of that was a driving force to say, "Hey, let's put this all under a legislative piece and let's put a little bit of teeth in this thing and so the districts do have a responsibility to report how these children are going to be doing, how they are mastering these Essential Skills, and when they are mastering them."

The mandate also demonstrated the state's attempt to create uniformity in classrooms across the state. The coordinator of the program declared that the Essential Skills

> really define what students ought to know and be able to do. Those are the documents the districts are to use for curriculum alignment. Those are the things that we are required to make sure are a part of our curriculum and are included in what we teach.

Data from interviews and participant observation in classrooms showed clearly that the compliance and uniformity expectations of New MDR were unfulfilled. Districts varied significantly in their capacities to accomplish the outcomes mandated. First, some districts were heavily invested in a traditional behaviorist pedagogy and had been rewarded over the years with high standardized and criterion-referenced test scores. Some districts with traditional curriculum and pedagogy might have had a desire to change, but lacked the money to retrain teachers or purchase new materials to address the cognitive-constructivist pedagogy and integrated curriculum that ASAP was designed to promote. Other districts were well on their way to this target, far ahead of others. Still others provided professional development activities that would spur teachers toward the goals of the state reform. Other than providing test forms and scoring workshops, the state paid little attention to professional development.

A key assumption of mandates is that the outcome is something that is required of all, regardless of their differing capacities (McDonnell & Elmore,

1987) and interpretations. In our study we examined the image of the classroom teacher with an eye on teachers' role in the student learning process. We found substantial variability in teachers' beliefs about themselves and about how students learn. The teachers we interviewed and observed also varied significantly in their classroom practices and their views of pupils and curriculum. The following two excerpts illustrate some of the differences we found:

Teacher A:

> **If I'm teaching facts and the things that the ITBS [Iowa Tests of Basic Skills] teaches, then I can open her up and pour it in—just open their little heads and pour it in.**
>
> **See the thing in our district is the review, the constant review, is almost as important as or more important than the initial teaching. One of our teachers used to say, "You've got to throw it against the wall enough times so it sticks in their minds."**

Teacher B:

> **Today we were doing something in our literature logs. It's just a follow-up to reading. Then they [students] had to do a kind of evaluating of the story, comparing, contrasting, that sort of thing. . . .**
>
> **You have to know what the Essential Skills are, and you have to be able to come up with ideas as to how to integrate those skills into the books, into the choices that the children make to read, with whole language. . . . It took me ten years to get to the point where I would be comfortable doing that. And I don't do it as much as a lot of them do it, even now.**

The difference between the two teachers illustrates variation in capacity to implement the mandated reform. Teacher A holds beliefs that are clearly less consistent with the cognitive-constructivist aims of ASAP than those of Teacher B. Furthermore, to progress from the beliefs about teaching held by Teacher A to the beliefs held by Teacher B is not simply an incremental adjustment but a fundamental change of thinking and practice.

To be consistent with cognitive-constructivist beliefs about learning and teaching, educational reform efforts directed toward instructional improvement

should acknowledge the challenges presented by such conceptual changes. Changes of this type are not simply brought about by the acquisition of new facts. Moreover, conceptual change typically occurs when there is dissatisfaction with the existing state of knowledge (Brown & Palincsar, 1989). Conceptual change is seldom achieved without attending to the beliefs of those who are the targets of change (teachers) and the conditions of the environments in which they function (schools). Our results show that such considerations were missing in the initial implementation of the ASAP reform. Instead, the Arizona Department of Education followed the behaviorist closed-system, believing that the testing mandate would create the means for achieving the desired ends of the reform.

By ignoring the process by which teachers must change from traditional curriculum and pedagogy to that which is consistent with the ASAP mandate, ADE violated the same cognitive-constructivist principles that it promoted for students. It treated teacher learning as a black box, as if adult, professional learning follows a different set of principles from children's learning. A respected body of thought (Candy, 1989; Dewey, 1933; Mezirow, 1991; Schon, 1983) suggests otherwise. To be consistent with the cognitive-constructivist model of teaching and learning, the reform effort would not be characterized as it was, by relegation of professional and curriculum development to the vagaries of the districts, by one-shot workshops and training sessions on test scoring, and by punishment through publication of poor results. Instead, a consistent reform would have first accepted *teacher* learning as a process of construction, and teachers as possessing diverse interpretations and prior knowledge structures. Second, recognizing the need for conceptual change and teacher learning in the context of classroom practice, the reform would have included sufficient time and resources such as mentoring, peer coaching, intensive seminars, and the like. Third, to encourage teachers to risk experimentation and failures in the short run, a political context where teachers feel safe to try new strategies would exist.

Our premise regarding the contradictions inherent in providing legitimacy to a behaviorist reform via a cognitive-constructivist theory was succinctly voiced by a member of one of the ASAP policy-making constituencies:

> **Teachers aren't going to become those kinds of instructors when they continue to be treated as empty vessels or deficient vehicles that need to be fixed.**

Development of teachers from a mastery model of content-deliverers to active constructors of knowledge and co-constructors of pupils' knowledge is not merely an incremental change but a fundamental shift in ways of thinking and acting. The framework that would support such a shift, that is, the quality and intensity of curriculum and professional development, the resources to support the development, and the time to incorporate and refine the changes, is not in place in Arizona, though it exists in some districts.

The Arizona Student Assessment Program and the National Test Movement

Efforts being made by the New Standards Project and the National Education Goals Panel parallel the thinking of the Arizona legislature and Department of Education. Two key assumptions hold true for Arizona's program and the national initiatives to reform education. First, it is assumed that by establishing higher standards and measuring progress toward them, teachers and students will be motivated to achieve higher levels of performance. Second, it is believed that through the use of a better test, an alternative assessment, the negative and unexpected consequences of high-stakes standardized testing will be minimized (Linn, 1993).

The national initiatives share the New MDR inconsistencies with ASAP: expecting that behaviorist, closed-system reform will direct schools toward higher standards and cognitive-constructivist aims. Furthermore, like Arizona's assessment plan and that of the United Kingdom (Torrance, 1993), the assessment reform attempts to fulfill two missions: specifying a direction of reform and mandating a high-stakes accountability device. So far, the argument that such models assume or ignore means to the end has been addressed in two ways (Porter, 1993). First the National Council on Education Standards and Testing (1989) has proposed that a national system of assessments be accompanied by a set of school delivery standards. These standards would specify what schools need to offer so that students can achieve the standards. Second, the New Standards Project specifies a "social contract" whereby members promise not to apply high stakes to assessment results until curriculum revisions are in place. Although these latter provisions suggest some relief from the old MDR, the legislative debate indicates that the constituency in support of high-stakes assessment is considerably stronger than the constituency

for school delivery standards. Likewise, the resources needed for testing are considerably less than what it would take to provide the texts, materials, and professional development necessary to bring schools up to the standards embedded in the reform. Viewing Arizona as a microcosm, we expect similar problems to befall the national measurement-driven reform efforts.

Alternative Policy Instruments

How can a human change initiative be effective if it ignores the differing capacities of those it ultimately wishes to influence? Reform mandated through measurement flies in the face of what cognitive scientists refer to as intentional learning, that is, learning desired and controlled by the learner. Individuals' construction of new knowledge depends strongly on their sense of being in charge of their learning (Bereiter & Scardamalia, 1989). When teachers do not feel in control of their own professional lives, they act passively, they become compliant and act automatically without reflecting on their own beliefs. Not only do mandates disregard the repertoires of practitioners; they may actually undermine teachers' capacity and motivation to change by stripping them of their autonomy (Mathison, 1991).

Mandates are not instruments that encourage practitioner inquiry or dissatisfaction with the status quo; on the contrary, they depend on coercion to create uniformity (McDonnell & Elmore, 1987). If one expects practitioners to change their practice, and for some this means challenging their current views of themselves and their students, an environment conducive to such change must be fostered. The teaching context (i.e., school environment, messages from administration, expectations of other teachers, and such) facilitates or detracts from the possibility of change. Combs (1991) described an alternative frame of reference as an open system. Following a personal growth model, an open system is founded on the belief that "things don't change people." Participants in an open system are viewed as active and responsible, goals are broad and not entirely predictable. Responsibility is shared and cooperative effort is valued. This model reflects much of the cognitive-constructivist perspective.

Capacity-building policy initiatives more accurately mirror the open system model. This type of initiative assumes that the desired capacity does not as yet exist and that investment is needed to facilitate its development. The expected effect is the enhancement of skill and competence, though these be returns that

are often uncertain, distant, and intangible. Central to capacity-building efforts is the recognition that, although short-term results may be limited, the longer term benefits are worth the wait (McDonnell & Elmore, 1987).

The Arizona Student Assessment Program embodies the contradictions inherent in the use of mandates to accomplish capacity-oriented goals. Its focus on compliance and control in effect undermines its potential to create the context necessary for educators to develop the level of competence desired by those who hope to reform education.

The voices of proponents of this new view of measurement-driven reform echo the call heard by a farmer in the film *Field of Dreams*: "If you build it, they will come" (Kinsella, 1982). However, unlike the power of a dream that brought the legends of baseball to a perfect field, bringing pedagogy to the level desired by reformers will take more than building the "perfect" test.

References

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

Baroody, A.J., & Ginsburg, H.P. (1990). Children's learning: A cognitive view. In R.B. Davis, C.A. Maher, & N. Noddings (Eds.), ***Constructivist views on the teaching and learning of mathematics. Journal for Research in Mathematics Education, Monograph Number 4,*** **51-64.**

Bereiter, C., & Scardamalia, M. (1989). Intentional learning as a goal of instruction. In L.B. Resnick (Ed.), ***Knowing, learning, and instruction: Essays in honor of Robert Glaser*** **(pp. 361-392). Hillsdale, NJ: Lawrence Erlbaum Associates.**

Bransford, J.D., & Vye, N.J. (1989). A perspective on cognitive research and its implications for instruction. In L.B. Resnick & L.E. Klopfer (Eds.), ***Toward the thinking curriculum: current cognitive research 1989 ASCD Yearbook.*** **Reston, VA: Association for Supervision and Curriculum Development.**

Brown, A.L., & Palincsar, A.S. (1989). Guided, cooperative learning and individual knowledge acquisition. In L.B. Resnick (Ed.) ***Knowing, learning, and instruction: Essays in honor of Robert Glaser*** **(pp. 393-451). Hillsdale, NJ: Lawrence Erlbaum Associates.**

Candy, P.C. (1989). Constructivism and the study of self-direction in adult learning. ***Studies in the Education of Adults, 21,*** **95-116.**

Cannell, J.J. (1987). ***Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average.*** **Daniels, WV: Friends for Education.**

Chi, M.T.H. (1978). Knowledge structures and memory development. In R. Siegler (Ed.), ***Children's thinking: What develops?*** **(pp. 73-96). Hillsdale, NJ: Lawrence Erlbaum Associates.**

Combs, A.W. (1991). ***The schools we need: New assumptions for educational reform.*** **Lanham, MD: University Press of America, Inc.**

Corbett, H.D., & Wilson, B.L. (1991). ***Testing, reform, and rebellion.*** **Norwood, NJ: Ablex.**

Darling-Hammond, L., & Wise, A.E. (1985). Beyond standardization: State standards and school improvement. ***The Elementary School Journal, 85,*** **315-336.**

Dewey, J. (1933). ***How we think.*** **Chicago, IL: Regnery.**

Dewey, J. (1963). ***Experience and education.*** **New York: MacMillan.**

Driver, R., Guesne, E., & Tiberghien, A. (1985). ***Children's ideas in science.*** **Philadelphia, PA: Open University Press.**

Erickson, F. (1986). Qualitative methods in research on teaching. In M.C. Wittrock (Ed.), ***Handbook of research on teaching*** **(pp. 119-161). NY: Macmillan.**

Fish, J. (1988). ***Responses to mandated testing.*** **Unpublished doctoral dissertation, University of California, Los Angeles.**

Gergen, K.J. (1985). The social constructionist movement in modern psychology. ***American Psychologist, 40*****(3), 266-275.**

Glaser, R. (1984). Education and thinking: the role of knowledge. ***American Psychologist, 39*****(2), 93-104.**

Glaser, R., & Bassok, M. (1989). Learning theory and the study of instruction. ***Annual Review of Psychology, 40,*** **631-666.**

Haas, N.S., Haladyna, T.M., & Nolen, S.B. (1989). ***Standardized testing in Arizona: Interviews and written comments from teachers and administrators*** **(Tech. Rep. 89-3). Phoenix: Arizona State University, West Campus.**

Haladyna, T.M., Nolen, S.B., & Haas, N.S. (1991). Raising standardized test scores and the origins of test score pollution. ***Educational Researcher, 20*****(5), 2-7.**

Hatch, A., & Freeman, E.B. (1988). Who's pushing whom? Stress and kindergarten. ***Phi Delta Kappan, 70,*** **145-47.**

Herman, J.L., & Golan, S. (1991). ***Effects of standardized testing on teachers and learning: Another look*** **(CSE Tech. Rep. No. 334). Los Angeles: University of California, Center for the Study of Evaluation.**

Holmes, C.T. (1989). Grade level retention effects: A meta-analysis of research studies. In L.A. Shepard & M.L. Smith (Eds.), ***Flunking grades: Research and policies on retention*** **(pp. 6-33). London: Falmer Press.**

Honig, B. (1987). How assessment can best serve teaching and learning. ***Assessment in the service of learning: Proceedings of the 1987 ETS Invitational Conference.*** **Princeton, NJ: Educational Testing Service.**

Kinsella, W.P. (1982). ***Shoeless Joe.*** **NY: Ballantine Books.**

Koretz, D.M., Linn, R.L., Dunbar, S.B., & Shepard, L.A. (1991). ***Effects of high-stakes testing on achievement: Preliminary findings about generalization across tests.*** **Paper presented at the annual meetings of the American**

Educational Research Association and the National Council on Measurement in Education, Chicago.

Koretz, D., Stecher, B., & Deibert, E. (1992). *The Vermont portfolio assessment program: Interim report on implementation and impact, 1991-92 school year* (CSE Tech. Rep. No. 350). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing/RAND.

Linn, R.L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*(1), 1-16.

Linn, R.L., Graue, M.E., & Sanders, N.M. (1990). Comparing state and district test results to national norms: The validity of claims that "everybody is above average." *Educational Measurement: Issues and Practices, 9*, 5-14.

Lipsky, M. (1980). *Street-level bureaucracy: Dilemmas of the individual in public services.* New York: Russell Sage Foundation.

Madaus, G.W., West, M.M., Harmon, M.C., Lomax, R.G., & Viator, K.A. (1992). *The influence of testing on teaching math and science in grades 4-12. Executive summary.* Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

Mathison, S.M. (1987). *The perceived effects of standardized testing on teaching and curriculum.* Unpublished dissertation, University of Illinois at Urbana-Champaign, Urbana.

Mathison, S.M. (1991). Implementing curricular change through state-mandated testing: Ethical issues. *Journal of Curriculum and Supervision, 6*(3), 201-12.

McDonnell, L.M., & Elmore, R.F. (1987). Getting the job done: Alternative policy instruments. *Educational Evaluation and Policy Analysis, 9*(2), 133-152.

McGill-Franzen, A., & Allington, R.L. (1993). Flunk 'em or get them classified: The contamination of primary grade accountability data. *Educational Researcher, 22*, 19-22.

McNeil, L.M. (1986). *Contradictions of control: School structure and school knowledge.* NY: Routledge & K. Paul.

Medina, N., & Neill, D.M. (1988). *Fallout from the testing explosion: How 100 million standardized exams undermine equity and excellence in America's public schools.* Cambridge, MA: National Center for Fair and Open Testing.

Mehrens, W.A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement: Issues and Practice, 8*, 14-22.

Mezirow, J. (1991). *Transformative dimensions of adult learning.* San Francisco: Jossey-Bass.

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

National Council on Education Standards and Testing. (1992). ***Raising standards for American education.*** **Washington, DC: Author.**

Noble, A.J. (1993). ***Measurement-driven reform: The interplay of educational policy and practice.*** **Unpublished dissertation. Arizona State University, Tempe.**

Nolen, S.B., Haladyna, T.M., & Haas, N.S. (1989) ***A survey of Arizona teachers and school administrators on the uses and effects of standardized achievement testing*** **(Tech. Rep. 89-2). Phoenix: Arizona State University, West Campus.**

Piaget, J. (1948/1974). ***To understand is to invent: The future of education.*** **New York: Viking.**

Popham, W.J. (1987). The merits of measurement-driven instruction. ***Phi Delta Kappan, 68,*** **679-682.**

Popham, W.J., Cruse, K.L., Rankin, S.C., Sandifer, P.D., & Williams, P.L. (1985). Measurement-driven instruction: It's on the road. ***Phi Delta Kappan, 66,*** **628-634.**

Porter, A. (1993). School delivery standards. ***Educational Researcher, 20*****(5), 24-30.**

Rein, M. (1976). ***Social science and public policy.*** **Middlesex, England: Penguin Books Ltd.**

Resnick, L.B. (1987). ***Education and learning to think.*** **Washington, DC: National Academy Press.**

Resnick, L.B. (Ed.). (1989a). ***Knowing, learning, and instruction: Essays in honor of Robert Glaser.*** **Hillsdale, NJ: Lawrence Erlbaum Associates.**

Resnick, L.B. (1989b). Toward the thinking curriculum: An overview. In L.B. Resnick & L.E. Klopfer (Eds.), ***Toward the thinking curriculum: Current cognitive research/1989 ASCD Yearbook*** **(pp. 1-18). Reston, VA: Association for Supervision and Curriculum Development.**

Resnick, L. B., & Resnick, D.P. (1989a). ***Assessing the thinking curriculum: New tools for educational reform.*** **Washington, DC: National Commission on Testing and Public Policy.**

Resnick, L.B. & Resnick, D.P. (1989b). Tests as standards of achievement in school. In ***Proceedings of the 1989 ETS Invitational Conference: The uses of standardized tests in American education*** **(pp. 63-80). Princeton, NJ: Educational Testing Service.**

Romberg, T.A., Zarinnia, A., & Williams, S.R. (1989). ***The influence of mandated testing on mathematics instruction: Grade 8 teachers' perceptions (monograph).*** **Madison: University of Wisconsin, National Center for Research in Mathematical Science Education.**

Schon, D.A. (1983). ***The reflective practitioner: How professionals think in action.*** **New York: Basic Books.**

Shepard, L.A. (1989). A review of research on kindergarten retention. In L.A. Shepard & M.L. Smith (Eds.), ***Flunking grades: Research and policies on retention*** **(pp. 64-78). London: Falmer Press.**

Shepard, L.A. (1991). Negative policies for dealing with diversity: When does assessment and diagnosis turn into sorting and segregation? In E. Hiebert (Ed.), ***Literacy for a diverse society: Perspectives, practices, and policies*** **(pp. 331-352). New York: Teachers College Press.**

Shepard, L.A., & Dougherty, K.C. (1991, April). ***Effects of high stakes testing on instruction.*** **Paper presented at the annual meetings of the American Educational Research Association and National Council on Measurement in Education, Chicago.**

Smith, M.L. (1991). Meanings of test preparation. ***American Educational Research Journal, 28,*** **521-542.**

Smith, M.L., Edelsky, C., Draper, K., Rottenberg, C., & Cherland, M. (1990). ***The role of testing in elementary schools*** **(CSE Tech. Rep. No. 321). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.**

Smith M.L., Noble, A.J., Heinecke, W., Cabay, M., Junker, S.C., & Saffron, Y. (1993). ***What happens when the test mandate changes: Final report.*** **Los Angeles: University of California, 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,*** **7-11.**

Stodolsky, S.S. (1988). ***The subject matters: Classroom activity in math and social studies.*** **Chicago: University of Chicago Press.**

Torrance, H. (1993). Combining measurement-driven instruction with authentic assessment: Some initial observations of National Assessment in England and Wales. ***Educational Evaluation and Policy Analysis, 15*****(1), 81-90.**