DOCUMENT RESUME

ED 381 550 TM 022 834

AUTHOR Noble, Audrey J.; Smith, Mary Lee
TITLE Measurement-Driven Reform: Research on Policy, Practice, Repercussion.
INSTITUTION Arizona State Univ., Tempe.; National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO CSE-TR-381
PUB DATE Aug 94
CONTRACT R117G10027
NOTE 32p.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Educational Assessment; Educational Change; Educational Policy; Elementary Secondary Education; Ideology; Measurement Techniques; *Policy Formation; Political Influences; *State Programs; *Test Construction; *Testing Programs; Test Use
IDENTIFIERS *Arizona Student Assessment Program; *Measurement Driven Instruction; Performance Based Evaluation; Reform Efforts

ABSTRACT

The Arizona Student Assessment Program (ASAP) epitomizes the principle on which measurement-driven reform is based, "You get what you assess." This policy study examines the ideologies and intentions of groups instrumental in the creation and implementation of a performance-based assessment reform. The study was conducted by interviewing members of the Arizona educational policy-shaping community and by examining documents and artifacts related to testing. It reveals both the ambiguities characteristic of the policy-making process and the dysfunctional side effects that evolve from the policy's disparities. Arizona's plan to reform its schools is still held captive by conflicting political forces and ideologies. ASAP appeals to many because of its ambiguity, but this same characteristic may undermine its capacity to bring about substantial change in educational practice. (Contains 30 references.) (Author/SLD)

CRESST

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

Measurement-Driven Reform: Research on Policy, Practice, Repercussion

CSE Technical Report 381

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

▸ UCLA Center for the Study of Evaluation

in collaboration with:

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

Measurement-Driven Reform: Research on Policy, Practice, Repercussion

CSE Technical Report 381

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

August 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 number R117G10027 and CFDA catalog number 84.117G, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

MEASUREMENT-DRIVEN REFORM: RESEARCH ON POLICY, PRACTICE, REPERCUSSION[1]

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

ABSTRACT

The Arizona Student Assessment Program (ASAP) epitomizes the principle on which measurement-driven reform is based, "You get what you assess." This policy study examines the ideologies and intentions of groups instrumental in the creation and implementation of a performance-based assessment reform. It reveals both the ambiguities characteristic of the policy-making process and the dysfunctional side effects that evolve from the policy's disparities.

INTRODUCTION

Educational reform initiatives over the past decade could be characterized as inconsistent and even antagonistic. While one group cries out for national standards and a national curriculum, another pleads the case for decentralization. Those who assert the need for improved neighborhood schools try to out-shout the proponents of school choice. The public demands for educational accountability have become entangled in the movement for site-based management. Funding debates rage. Will vouchers solve the problem? Or are equitable funding formulas the answer? Advocates of world class standards argue among themselves about whether schools should emphasize basic skills or critical thinking. These contrary trends also surface in the debates over the role of assessment. Among those who argue that testing reforms will improve schools, there is controversy about the form of testing that should be used. Traditionalists favor norm-referenced, standardized tests such as the California Achievement Test and the Iowa Tests of Basic Skills. New

[1] This work was also reported in a paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana, April 7, 1994.

voices have argued for performance exams that require students to understand and apply higher level thinking skills. These different positions share the assumption that when the state or federal government requires that the school measure pupil achievement, teachers and pupils will try harder to raise achievement, and thus, education will improve.

The legislative passage of the testing mandate in May 1990 demonstrated Arizona's commitment to top-down reform and its belief that assessment can leverage educational change. Arizona Revised Statute 15-741 directed the State Board of Education to adopt and implement a performance-based assessment plan aligned with the state's curriculum, the Essential Skills. The Arizona Student Assessment Program (ASAP) was the Arizona Department of Education's (ADE) response to the legislation. The plan straddled the faultline of the testing controversy, incorporating both standardized and performance-based assessments.

The purpose of the research was to portray how this reform was conceived, negotiated, and implemented, as well as to document initial responses to it. Particular attention was paid to the images, ideologies, values, and goals of those involved in the development and implementation of the Arizona Student Assessment Program.

LITERATURE REVIEW

The Arizona Student Assessment Program (ASAP) is based on the notion of measurement-driven reform. Popham outlined the traditional concept of measurement-driven instruction (MDI) stating that 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. Arizona policy makers relied on this assumption in their creation and implementation of their test-driven reform effort.

However, some have argued that the linkage is not so direct. According to Linn (1993, p. 3), "considerable caution is needed in using achievement test results to draw inferences about the quality of education." 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 and colleagues (Smith, Edelsky, Draper, Rottenberg, & Cherland, 1990; Smith & Rottenberg, 1991) found that schools neglected topics that the mandated tests failed to cover, such as science, social studies, and writing. 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.

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.

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.

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. Torrance (1993) described the United Kingdom's efforts to implement its National Curriculum through the use of a National Assessment. He 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 Student Assessment Program is this state's interpretation of the principle "you get what you assess." The findings of our study revealed that the conceptualization and implementation of measurement-driven reform is not as uncomplicated as its advocates allege.

RESEARCH DESIGN AND METHOD

Theoretical Framework

The work of Rein (1976) provided the theoretical framework of the study. First, he contended that the policy researcher should treat the purpose of social policy as unresolved. Palumbo and Calista (1990) agreed that the definition of a social problem and the subsequent policy design are products of conflict in that they result from political bargaining and compromise. Hence, the relationship between the intention of policy and the response of practice is seldom

straightforward. From this notion, the present study focused on identifying the various images, values, and purposes of the Arizona measurement-driven reform effort as potentially variable and conflicting.

Second, Rein argued that the questions directing policy analysis should challenge established patterns since policy grows out of political negotiation. When the analysis runs counter to prevailing trends, the contradictions and vulnerabilities of the policy can be discovered. Questioning orthodoxy avails the policy analyst of important issues that may have been left unattended or ignored. Rein believed that the analyst must approach social policy critically, realizing that all interventions may be regarded as ways of doing one thing and, at the same time, forsaking some other action. Therefore, the study of any policy in practice needs to examine the consequences of what may appear as widely shared principles. In response, the study explored the assumptions underlying the mandate and their implications. Following Finch (1986), the purpose of policy study was twofold: (a) to describe and understand the real effects of policies, and (b) to compare the assumptions upon which policies were based with social experience.

Finally, Rein argued for the necessity of considering the political reception of a policy study. Different types of knowledge tend to be used in different arenas. Lindblom (1980) suggested that instead of seeking to make recommendations, the research should be tailored to challenge policy makers' and practitioners' ways of thinking. In lieu of offering solutions to problems, effective policy research fosters conceptual reorientation of the issues and concerns that policies address. An outcome of the research was to reveal the complexity of the change process implied by measurement-driven reform.

Methods

Following Majchrzak's (1984) argument that policy research must attempt to study the multidimensional nature of the problem, this study employed multiple methods. The policy study was conducted by interviewing members of the policy-shaping community and by examining documents and artifacts of the testing policy. Interviews with key policy makers and stakeholders, representatives of groups central to the creation and development of ASAP, along with analysis of documents, provided evidence at the macrolevel. Those interviewed included a key Arizona state senator, officials in the Arizona

Education Association, a university professor who served as an advisor to the program, involved members of the Arizona Department of Education (specifically, ASAP personnel), and the State Superintendent of Public Instruction. Content analysis of the legislative mandate as well as documents related to ASAP was performed. The key policy documents included Arizona statutes and Senate bills, minutes from legislative hearings, the ASAP User's Guide, an ADE newsletter for teachers, and copies of the ASAP practice tests, including ADE directives to districts and teachers. Each document was examined to determine how policy makers and stakeholders defined the goals of the mandate, specifically as it would affect students, teachers, curriculum, assessment, and schools as organizations. Examination of policy issues through documents and semistructured interviews was supplemented by data gathered at workshops sponsored by the Arizona Department of Education. These included regional meetings and school-based workshops to familiarize school personnel with the purpose of and use of ASAP practice tests, along with statewide workshops to train teachers in the use of the performance test scoring rubrics. These activities provided a source of data similar to documents, nonreactive to the researcher's presence.

The examination of policy issues was directed toward constructing a narrative of the conception and implementation of the test mandate. An intent of the narrative was to produce a chronological presentation of the events surrounding the Arizona Student Assessment Program. Moreover, a goal of the analysis of the policy issues was to illuminate the assumptions of those who were most influential in the creation and implementation of a measurement-driven plan for educational reform. Through use of constant comparative methods of analysis (Strauss, 1987), the researchers examined the constructs held by each stakeholder relating to the ideals, assumptions, and perceived effects of the mandated testing program.

The process of data coding began early in the study and continued throughout the data collection. Assertions, based on the coding of data and discovered categories, were then constructed. Analytic narratives or vignettes, along with quotes from the data, instantiated the assertions. All interviews were recorded and verbatim transcripts of the tapes were produced. HyperQual, a computerized qualitative data analysis program, was used to manage the

various data sources, to organize research memos, and to code and sort data during the analysis process.

Numerous sources of data provided varied ways to look at the research question. Documents and interviews revealed the ideals of the policy itself. Throughout the study, the processes of data collection and analysis were iterative. A major portion of the data analysis occurred while the data were generated. Although these processes were interactive, for purposes of clarity, they are presented in a linear manner.

FINDINGS

The Political Venue

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 amounts of time in preparation for them (Haas, Haladyna, & Nolen, 1989). We identified two primary constituencies supporting a change in the test mandate. Each group had its own ideologies and interests. One constituency was 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. Its stakeholders combined forces with those who opposed the test because of its deleterious effects on students, teachers, curriculum, and instruction. Thus, a group most interested in accountability and outcomes joined with one devoted to instructional improvement and process. The alliance resulted in legislative action to change the test mandate.

This alliance was the first instance of efforts to merge diverse ideologies and purposes. Those members of the outcomes-oriented group, unhappy with the use of norm-referenced tests as an accurate accountability measure, argued that "they don't measure what our teachers teach." Its members were also

discontented with what they perceived as the lack of teacher or school attention to the state's curriculum framework and wanted to force districts to align their curriculum to the Arizona Essential Skills. A policy maker who represented this group discussed the concern.

> The Department of Education and the state legislature had learned that simply having the state Essential Skills in no way was serving as a catalyst for districts to align their curricula. An assessment program, they felt, was a way to do that. It would compel districts to finally do what they were supposed to have been doing for years.

From this group's viewpoint, the Essential Skills represented what schools should target as outcomes of instruction. Their goal was to make schools more accountable for student achievement.

The other partner in the alliance, the process-oriented group, hoped to change the kind of pedagogy that schools should adopt. While the outcomes group focused on ends, the process group focused on means. They hoped that the performance assessment would encourage teachers to adopt a holistic, constructivist pedagogy that in turn would result in more meaningful learning for students.

> We really want to change curriculum so that the students are vitally engaged in their learning process and beginning to create their own knowledge.

As a part of this effort, they also acknowledged that the role of the teacher would need to change.

> The teacher as the deliverer of information, the teacher as 'I talk and you take notes,' that would be a thing of the past . . .

Although the ideals of the two groups were distinctly different, their political interests converged in the creation of the Arizona Student Assessment Program (ASAP).

The legislative effort brought together stakeholders who traditionally were at odds with each other. Republican and Democratic members of the Arizona legislature participated. The leadership of the Arizona Education Association and the Arizona Superintendents Association collaborated. The Arizona Department of Education worked with the Arizona School Board Association. Educational researchers from local universities participated in the discussions.

This political compromise was described by one school administrator as an indication that

> the system is capable of making significant reform and change, and that the system can be adaptive and responsive. (Baracy, 1992)

Seen from another vantage point, this compact was a political attempt to meld conflicting beliefs and intentions. Following Rein (1976), one can assume that the objectives of policy are multiple, ambiguous, and conflicting. Ambiguity is an essential element of political negotiation enabling agreement among competing viewpoints. According to Lindblom (1980), as a result of this untidy process of policy making, political compromise often results in a policy on which neither side had planned. Such was the case with the Arizona Student Assessment Program.

The alliance of interests based on conflicting images and intentions carried over from the legislation into the ASAP implementation plan. At each level, implementation and practice, individuals responded to conflicting messages, accepting those which were most understandable and screening out the others. Rein (1976) concurred with Lindblom that ambiguous and inconsistent legislation shifts the arena of decision to a lower level, to the level of implementation. The subsequent analysis illuminates these conflicting ideologies and how the Arizona Department of Education interpreted them.

The Ideological Inconsistencies

The ideals presented by the alliance corresponded to two conceptually different views of learning, behaviorist and constructivist. The outcomes-oriented group proposed a behaviorist, traditional learning view while those who valued process embraced constructivist learning theory. Each of these perspectives shaped the Arizona Student Assessment Program. These alternative paradigms fostered a cacophony of messages sent by policy makers to those responsible for implementing the program.

Inconsistency #1: Policy makers' definitions of "learning" were incoherent.

The Arizona Student Assessment Program was an outgrowth of the state Goals for Educational Excellence legislation. One of these goals was to increase

the level of achievement of all students in the public schools. As policy makers spoke of this goal and ASAP's role in accomplishing it, they revealed their beliefs about how children learn.

Those who embraced a behaviorist view of learning talked of the importance of outcomes and the mastery of the Essential Skills, the state's curriculum framework.

> As soon as the statute went into effect, we immediately got nine more Essential Skills documents. The statute says that they must be provided in all nine subject areas required by law. So there you have the Essential Skills documents—what kids need to know and do. (ASAP Unit coordinator)

This outcomes-oriented group affirmed that the Essential Skills was the body of knowledge that students needed to master during their years in school. The ASAP tests served as indicators of student achievement. They defined learning as the accumulation of skills over a period of time. These voices told teachers that learning was about the mastery of skills, the achievement of outcomes; competence was revealed in the scores their students achieved on tests.

However, while supporters of outcomes-oriented reform spread their message, another voice spoke out as well—those who valued the process of learning, those who defined learning from a constructivist viewpoint. Their declarations came from very different beliefs. They described learning not as an outcome, but as an interaction between the teacher and the student, one that fostered conceptual understanding and growth. In the 1992 version of the Arizona Essential Skills for Mathematics document, those who supported this view of learning attempted to alter behaviorist interpretations of earlier versions of the document.

> The 1987 version of the *Arizona Essential Skills for Mathematics* has been interpreted by some as a checklist of isolated skills. The intent of the [Mathematics Essential Skills] document was far more comprehensive . . . Mathematics must be fully explored within the context of the real world. Although content is important, it is one's ability to problem solve that ultimately determines the outcomes of one's encounters with life. (1992 version of Arizona Essential Skills for Mathematics)

The messages sent by those who viewed learning traditionally clashed with those who saw learning through a constructivist lens. This conflict was

communicated to practitioners through numerous avenues (e.g., workshops, the Department of Education's newsletter, scoring training sessions, Essential Skills documents, media releases, the ASAP User's Guide). Unfortunately, the messages about learning were not the only incoherent pronouncements.

Inconsistency #2: Policy makers held dissonant expectations of teachers.

The inharmonious voices of policy makers were heard in relation to how they defined a "good" teacher. While behaviorists declared that "good" teachers were those who delivered instruction according to the Essential Skills, constructivists spoke of teachers facilitating learning. Those with behaviorist leanings articulated that the teachers' responsibility was to teach skills so that students could perform the Essential Skills.

> People [teachers] need to know that the Essential Skills are the framework that they have to work around or within. I mean they have to. Whatever they're going to put in place has to get the kids ready at the end of third, eighth and twelfth grade to pass those examinations. That's what they're working for, nothing else. (State Superintendent of Public Instruction)

According to this group, the authority, the source of all important knowledge was the Essential Skills. The teacher's job, according to those who valued outcomes, was to transmit the information from the curriculum to the students. On the other hand, the proponents of constructivist learning described the teacher's role as one who organizes resources, coaches students and learns collaboratively with other teachers.

> The role of the teacher would be to develop pathways to resources. I see the teacher as coach. (past ASAP coordinator)

Proponents of learning as a process encouraged teachers to value their own growth as professionals as a complement of their students' learning. In their minds, the source of knowledge was what occurred in the interaction among the students and the teacher. The definitions of knowledge, according to these two groups, were further revealed in their interpretations of the role of curriculum.

Inconsistency #3: Policy makers clashed regarding the role of curriculum.

Curriculum played a key role in the Arizona education reform effort. All policy makers saw curriculum as instrumental in effecting change. Yet, their messages clashed in regard to the role of curriculum. Those who valued outcomes described the curriculum as an end, in and of itself. They embraced the state curriculum frameworks as representing the high standards of achievement that should be reached by all students in Arizona. The Essential Skills encompassed what all students should learn and what all teachers should teach.

> The Essential Skills documents really define what students ought to know and be able to do . . . Those are the things that we are required to make part of our curriculum and are included in what we teach . . . The Essential Skills are the core of what we'd expect Arizona students to learn. (ASAP coordinator)

Upon closer scrutiny of these documents, the number of skills in two of the curriculum areas alone, language arts and mathematics, revealed what the behaviorists defined as a "core" of instruction (Table 1).

Regarding the role of curriculum, the voices of those supporting holistic instruction also spoke out. In their view, curriculum was only a means to promote learning. They encouraged teachers to view content as a "dynamic process" not a "static discipline." They advanced the ideal that curriculum was a multifaceted integration of content and process, one in which the student's

Table 1

Number of Arizona Essential Skills by Curriculum Area and Grade

Grade levels	Language Arts[a]	Mathematics	Totals
K-3	62	37	99
4-8	73	77	150
9-12	102	100	202

[a] Language Arts includes skills in reading, writing, speaking, listening, and language concepts.

development was central. They continuously asserted that the process implied by the Essential Skills documents represented the "best we know" about learning and instruction.

> The language arts Essential Skills really are a total curriculum, and they do follow the best we know about research. When we say that they're coordinated with the cognitive statements, it means that they follow the stages of development. (ASAP coordinator)

Whether curriculum would be seen as a means or an end was very much dependent on the role assessment played in the reform. The cacophony was most shrill when policy makers spoke out about testing and what they hoped it would accomplish.

Inconsistency #4: Policy makers alleged that a single performance assessment could fulfill the dual purposes of instructional improvement and accountability.

Early in the reform initiative, the role of assessment was a source of controversy. A state senator who headed the Senate Education Committee that drafted the original Goals for Educational Excellence legislation shared her memories of how assessment had become an issue.

> The truth of the matter is, I didn't think about testing in the formulation of the Bill. If I did, it would have had a short fuse, and the Bill would have exploded in the process . . . when you are looking at wanting to have significantly better education for students throughout their elementary and high school experience, you realize that [standardized] testing wasn't showing you what you need.

When asked about this statement in a later interview, she explained her initial reticence to introduce a new form of testing.

> Most legislators felt that the mandatory, you know, the norm-referenced and criterion-referenced testing was the answer.

However, she, as a former elementary teacher, believed there were limitations to the state testing program. As she spoke, she recounted the questions she confronted as the legislative group struggled with the role of assessment in education reform.

> How are you going to get to the issue of higher order thinking skills? How are you going to get to the issue of the capacity to write? How are you going to get to the issue of comprehension and broad understanding and knowledge as to whether or not you are making progress and preparing the students for life as it is today?

She spoke of testing as a means to encourage teachers to teach more conceptually, focusing on higher order thinking skills, comprehension, and the like. However, in her next statement she turned to the accountability function of testing, that is, "how well our schools are educating our children."

> In other words you just came to the conclusion that using only a statewide norm-referenced test, that some people were reticent to give up, that it wasn't giving us the information we needed about how well our schools are doing in educating our children. And it just came as a focus point. Very clearly as a focus point. You either made progress on that point, or you couldn't move onto all the other issues that needed to be done . . .

The questions she asked herself foreshadowed the conflicts of purpose regarding assessment. Even as she spoke of using testing to gain a better understanding of higher order thinking, she also asked herself accountability questions. The confusion of purposes, whether testing should serve to improve instruction or evaluate it, had begun before the legislation had even been drafted. She also alluded to the allegiance some policy makers had to norm- a l criterion-referenced tests, believing them to be bett r suited to accountability, that is, to reveal specifically how well schools were fulfilling their responsibilities and to prod them to do better by publicizing poor performance. She referred to another legislator who had been active earlier in the state accountability movement.

> And J. H. who was the mother of testing in about '81. [Testing] was the legislators' way of saying 'educators, you're going to teach kids and we're going to know what they learn.'

New evidence established that norm-referenced tests were not measuring the state-mandated curriculum frameworks. The research (Haas et al., 1989) found that only 26% of the Essential Skills were measured by the Iowa Tests of Basic Skills and the Tests of Academic Proficiency. Therefore, those in the accountability camp acknowledged the need for an alternative assessment as a better measure of accountability to coerce districts to align their curricula with the Essential Skills.

> It was a matter of here we have the Essential Skills and I think there was ample evidence that many school districts weren't getting at that . . . Teachers were still teaching what they were teaching. They weren't focusing on those Essential Skills. And I think part of that was a driving force to put this all under a legislative piece and put a little bit of teeth into this thing. (Pupil Achievement Testing Unit coordinator)

The proponents of accountability saw the performance tests as a means to *force* districts to align their curriculum which, in turn, would result in better teaching.

> Because we had those Essential Skills for a number of years. But we weren't testing to those. We were testing the Iowa and people were teaching the Iowa and just ignoring the Essential Skills.
>
> (Interviewer): So what you're saying is whatever the high stakes are attached to is what people will teach?
>
> That's exactly right. (State Superintendent of Public Instruction)

Meanwhile, the advocates of constructivist learning theory believed that performance testing could be a means of instructional improvement. They saw these new test forms as ways to *encourage* teachers to teach more holistically, focusing on higher order thinking skills.

> And our thought was that if we developed a system of assessments that was based on quality instructional methods that teachers inherently would work themselves into better instruction by using and understanding the assessments.

Those who spoke of performance assessment as a vehicle of instructional improvement emphasized the integration of assessment and teaching.

> This is something that is embedded in the instructional process, day after day. The idea is for you to use them so that when the kids see this assessment in March, they'll say 'oh, this is something that we've been doing all the time.' The idea is that it becomes a part of the teaching and learning process.

The discord among the ideals of the policy makers regarding learning, teachers, curriculum, and assessment reverberated throughout the first year of implementation. A political alliance that brought together groups with conflicting values and beliefs created the Arizona Student Assessment Program. As the ideals were shared with educators across the state, most saw value in the program. The implementation plan appealed to many practitioners because of its ambiguity. Proponents of outcomes-based education saw ASAP as supporting

their ideals as it advanced standards and student mastery of skills. Simultaneously, advocates of whole language instruction thought ASAP fostered constructivist learning theory, emphasizing process.

Policy makers seemed unaware of the inharmonious messages that were being sent to practitioners. The confusion was magnified by the fact that the contradictory messages were not only sent by members of separate groups. The contradictions were often voiced by one individual. The empirical data used to illustrate the previously discussed assertions were not always spoken by members of the one group or the other. Policy makers at one moment would proclaim the value of the mastery of skills and the next would advocate integrated, conceptual learning.

While examining the first year of implementation of this policy it is easy to understand Weatherley and Lipsky's (1978) contention that the heavy overload of demands and expectations resulting from new policies means that street-level bureaucrats are essentially free to develop their own coping devices. The overload of messages from ASAP was not only heavy but conflicting. The inconsistencies apparent in the policy ideals sent conflicting messages to those who needed to implement the program. The next section illustrates how incongruous ideologies manifested themselves in the implementation plan developed by the Arizona Department of Education.

Repercussions of Incongruous Policy

Inconsistency #5: The implementation plan of the Arizona Student Assessment Program is a dysfunctional side effect of a policy built on contradictory ideals.

As an instance of measurement-driven reform, ASAP is internally inconsistent. This inconsistency has evolved as those responsible for implementing the policy tried to make sense of the numerous, conflicting messages generated by themselves and others. Similar to any learner confronted with new information, these individuals brought their prior knowledge and beliefs about schools, teaching, and testing to the implementation arena. As they created this new program of reform, based on a relatively unknown form of assessment and a new learning theory, their values, their past

experiences, and their political intentions came into play. An examination of the current status of the program revealed how the officials of the Arizona Department of Education interpreted the policy ideals according to their own inclinations and developed implementation strategies in kind.

ASAP is defined by the Department of Education as "a comprehensive program to improve teaching, learning, and assessment." However, as the plan of implementation evolved, one group of individuals became responsible for the improvement of teaching and learning while another group absorbed the responsibility for assessment. The stage was set for the process-product, instructional improvement-accountability debate. The division also foreshadowed the battle as to whether ASAP was to be a low- or high-stakes program.

For purposes of clarity, the two units responsible for the implementation, the ASAP Unit and the Pupil Achievement Testing Unit, are discussed separately even though their activities occurred simultaneously.

The ASAP Unit

The wavering voice of leadership. Since the inception of the performance-based assessment program in 1990, the leadership of the ASAP Unit has passed through four individuals. The variation in leadership and direction is illustrated by their comments regarding the goals of ASAP.

Leader #1 (based on her personal research during pilot year of the program):

> The ASAP as policy has been called the most profound incentive for change there has been in Arizona. Policy obviously cannot mandate what matters, but perhaps it can establish conditions for what matters.

Later during her term as director of the ASAP Unit she stated:

> This program calls for a much-needed improvement in Arizona's assessment system. The norm-referenced standardized test will never disappear in Arizona. The only hope is that its effect can be diminished and other, more authentic means of assessment valued. ASAP calls for a total revamping that is sure to have a profound effect on curriculum and instruction. These authentic assessments in reading, writing, and mathematics mirror the best we know about instruction in those subjects.

Leader #2 (during her interview two months after she had retired from the position as ASAP director):

> I think in the Goals for Educational Excellence, we had to say 'hey we need to take another look at assessing what we're doing.' . . . You know [if you] drill and kill on certain facts and information long enough, you're going to raise test scores. But are we teaching students how to think? Are we teaching critical thinking skills? Do we want a quick fix and raise grades or do we really want to change curriculum so that the students are vitally engaged in their learning process and beginning to create their own knowledge? So assessment then was looked at as the tool—if you change the way you assess, you're going to change curriculum, you're going to change the way instruction happens in the classroom and I think those were some of the goals of the program.

Leader #3 (comments from a workshop presented to Arizona teachers):

> We believe that the Arizona Student Assessment Program has the potential for bringing about major changes in Arizona—changes in what we do in the Department, changes in what happens in classrooms, in the organizations of schools, in teacher preparation programs, and most importantly, in addressing the following question. How can we educate every student who comes to our school—every student—and still take the whole to a higher level? And in a nutshell, that describes the purpose for ASAP. Every child has to leave our school being successful at a higher level than they've ever had to before. It is critical. No one is expendable.

Leader #4:

> ASAP is a systemic change. The Essential Skills have been in place for years. And no one has responded to them until we started assessing them. I realized that this is the most powerful program. And it's a systemic change program. And that's the key to it . . . One of things I've been bothered about all the years I've been in education is that we never really decided what we think is important for kids to learn. Okay, the Essential Skills really say that all kids should do this.

Each of these directors interpreted the purpose of ASAP somewhat differently. The first director hoped for curriculum and instruction to be more aligned with cognitive learning theory. The second director followed this lead but reinforced the value of fostering higher order thinking skills. Leaders 1 and 2 supported the values of the constructivist viewpoint, valuing process over product. However, the next leader shifted the program's attention to the outcomes of learning, emphasizing student competence. The final and current leadership directed the program toward the state curriculum frameworks, the

Essential Skills, as the desired outcomes of instruction. Although the differences appeared subtle, each of these leaders' interpretations led the implementation of the program in different directions.

Function: Instructional improvement. The ASAP Unit was primarily responsible for the instructional improvement focus of the implementation plan. The plan took two forms: the distribution of alternate tests and the provision of scoring training workshops.

The primary function of this unit was to provide teachers with alternate forms of the performance tests in reading, writing, and mathematics. Over 200 of these forms had been distributed to schools prior to and during the 1992-93 school year, the first year of implementation. The intent of these forms (Forms A, B, and C) was to "facilitate teaching and learning at all grade levels . . . Because the assessments reflect good teaching practices, teachers may use them as instructional units or models for developing effective instructional strategies." In addition, an intention of the ASAP Unit was that teachers could administer these test forms to students, on an ongoing basis, when they felt that students were competent in the skill(s) that the particular form assessed. Testing would be integrated with instruction. They described this as a "one-to-one match," testing matching curriculum.

A second means by which the ASAP Unit planned to improve instruction was through scoring workshops. They conducted these training sessions across the state to prepare teachers to score the performance assessments according to a generic rubric. The assumption underlying this training effort was that if teachers knew how to score assessments using the 5-point rubric, they would become better teachers. One director validated this assumption at an orientation workshop.

> The greatest advantage is when one understands the scoring. What we've discovered when we scored the pilot As, we got feedback from the scorers and every one of them, uh, just about everyone, said 'I know how we teach this now. Now that I know the standard I'm looking for, I can teach it.' (current ASAP coordinator--Leader #4)

However, limited finances restricted the Department of Education to the number of workshops it could conduct, so the expectation was that teachers trained in scoring would teach others in their schools.

> Hopefully, you were informed that there is an expectation that you will now be going to go back and train others back in your districts or the surrounding districts. We need to get more and more teams of people trained so that they can go back to school and train others and then we can get all 37,000 teachers in the State of Arizona trained to use the rubrics for scoring assessments. (ASAP coordinator–Leader #3)

These methods of instructional improvement—test forms and scoring workshops—demonstrated an inherent contradiction between the ideals which these individuals verbally promoted and the strategies used to implement their plan. According to one ideal, the program promoted constructivist assumptions about psychology and learning.

> Performance-based assessment encourages a learning environment where students learn for greater understanding . . . they work cooperatively, analyze and discuss their thinking processes. (ADE ASAP brochure)

However, logistically, the implementation was built on behaviorist assumptions about reforming schools and teaching teachers. The learning principles promoted for students were ignored as they applied to teachers. Teachers' learning was reduced to receiving hundreds of unfamiliar test forms and being trained in scoring. Contrary to the ideal where instruction enables students to make meaningful connections between what they already know and new information, the intention of the ASAP Unit was to standardize teachers' learning. The current director spoke of the most recent training efforts towards this end.

> Another thing that we'll be grilling on this year is that we will be providing some more materials. We want to provide the scoring booklet for you so that at third grade you can see what a two looks like, what a three looks like, using some of the As, Bs and Cs, so you can actually see how these are scored. [This is] so that you can see the development from a one to a two to a three to a four. (current ASAP coordinator)

The ASAP Unit's implementation plan contradicted what its personnel professed to promote about teaching and learning. An excerpt from a descriptive brochure of ASAP expressed the state's concern about why Arizona chose a new way to assess.

> Old ways of teaching and testing have created students who are, all too often, *passive recipients* of information instead of *active learners*. (their emphasis)

These same beliefs were obviously not held for teachers. The contradictions inherent in providing legitimacy to a behaviorist reform via a constructivist theory was succinctly voiced by a member of one of the ASAP policymaking 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. (Arizona Education Association representative)

While the ASAP Unit vacillated on its interpretation of ASAP as vehicle of instructional improvement, the Pupil Achievement Testing Unit maintained its position as an "auditor" of district accountability.

The Pupil Achievement Testing Unit

The steady voice of leadership. Prior to the change in the testing mandate, all students in Arizona were tested each spring on either the Iowa Tests of Basic Skills or the Tests of Academic Proficiency. This norm-referenced testing was administered by the Pupil Achievement Testing Unit. Along with the coordination of the test administration and scoring, the unit's director would prepare a statewide report, required under legislative mandate, for the State Superintendent of Public Instruction. Each summer this report appeared in major newspapers across the state. School districts were ranked according to their students' scores on the tests. These reports also frequently appeared in realtors' advertisements as attempts to lure prospective buyers into particular school districts. When the test mandate expanded the assessment program to include performance assessment, the required standardized testing was moved to the fall and limited to three grade levels. The state-required performance tests were to be administered in the spring. The stated intent was to diminish the direct accountability function of norm-referenced testing. The director of the Pupil Achievement Testing Unit said the following about the change:

> Many districts don't have a good strong sense about how to use the norm-referenced test. Now [with the fall administration] there is a way to use it relative to curriculum and instruction. If they in fact felt that their kids needed so many lessons on certain kinds of spelling or word usage or whatever, and they take this very early and they find that they did very well on it, then they can refocus that instruction to some other ways. (Pupil Achievement Testing coordinator)

Another ADE official spoke of how decisions about student progress would be made with this change in testing.

> It [standardized test] tells you how the students of this school did against a national norm. This time we're going to compare how the students in that school did against a set of learning expectations. (State Superintendent of Public Instruction)

The leadership of this unit was clear in its view of the purposes of the new testing program. The norm-referenced tests would serve a diagnostic function and be used to meet federal standards for funded programs, such as Chapter 1. The performance tests were to measure students against competency standards based on the Essential Skills. Much of the responsibility for accountability reporting, under the new test mandate, would be on the individual school districts.

> Built all into this also is the accountability at the district level. The districts probably have the majority of the onus of responsibility now. Up until this time, the district always had the onus of responsibility to report results. But they will [provide] a district Completion Report. Basically, what they have to do is to report the percent of those kids that are mastering [the Essential Skills]. (Pupil Achievement Testing coordinator)

Although the "onus of responsibility" was left to the districts, the Department of Education retained dominion over accountability. The director ended his comments with "And, we will also be reporting the same thing."

Function: Accountability. Although the original plan for performance assessment did not include a point-in-time assessment, the Department of Education chose to retain a key element of its accountability function. The Pupil Achievement Testing Unit resumed its role, this time as an "auditor" of districts' progress. Form D, the "secured form" of the performance test, was created to serve that function. The proposed intent of Form D was to verify or "audit" the Completion Reports submitted by the school districts.

> We developed the assessment tool to confirm or negate some of the kinds of things that the districts were doing. And that's what [Form] D is about. (Pupil Achievement Testing coordinator)

Form D, although a performance-based test in appearance, took on many of the characteristics of the previous standardized testing. In contrast to the

Forms A, B, and C, distributed by the ASAP Unit, Form D, albeit performance-based, was a "standardized" 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," the ADE constructed scoring criteria, the generic rubrics. In lieu of having the assessments scored by trained Arizona teachers, over 75% of the tests were scored by Measurement, Inc., a subsidiary of the testing company that created the test forms. Results of the 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," were illustrated in a frequency distribution of student scores and by the 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.

The Pupil Achievement Testing Unit of ASAP had successfully maintained its function of accountability through the creation and administration of Form D. In contrast, the ASAP Unit personnel continued to declare that the intent of ASAP was instructional improvement. The ultimate contradiction inherent in the policy ideals regarded the issue of stakes. The disparity substantiated Madaus' (1988) claim that the level of stakes is the extent to which individuals "perceive" test performance to be used to make important decisions that immediately affect them. This suggests that stakes are a characteristic of the reactions of individuals or groups, rather than some inherent quality of the test or the testing policy itself.

Inconsistency #6: ASAP is both a high- and low-stakes assessment.

During the first year of implementation, prior to the administration of Form D, some teachers saw value in ASAP as an instructional improvement vehicle. Teachers from different schools made these comments.

> I will say that I got more out of giving my own students the ASAP and grading it myself but that's going to only help me with my class. They shouldn't use it to evaluate your class in our school or in our state.

> The ASAP is all process oriented so we have to go with process stuff . . . more critical thinking and prob'em solving in math now rather than calculations. So it's changed, so because we're teaching that way, the test is geared that way.

During an interview with the current ASAP Unit coordinator, I shared the interpretations of some practitioners of the purpose of ASAP.

> Interviewer: During a lot of the interviews with teachers and school administrators, it seems that many of them felt that the true intention of the performance assessment was actually to help them improve instruction in their classrooms.
>
> Coordinator: They felt that? Oh, that's great! That's how I view it.

Throughout the first year of implementation the ASAP Unit continued in its efforts to promote the value of performance assessment as instructional improvement. However, concurrently the Pupil Achievement Testing Unit pursued its auditing function. An ADE official described his perceptions of the high-stakes nature of ASAP.

> What's important to me now is that ASAP after the first administration of Form D in 1993 is now being reported out by school, district and the state. Now, that's where the action is. That's where the stakes are. (Associate Superintendent of Educational Services)

After the administration of Form D, the Pupil Achievement Testing Unit pursued its function just as it had in prior years with the Iowa test scores. 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.

The most recent (January 1994) accountability action was taken by the Arizona State Board of Education. Upon the recommendation of the State Superintendent of Public Instruction, the Board passed a resolution that, beginning with the ninth-grade class of 1996, high school graduation will be based on competency as determined by student performance on the ASAP tests. Following the state hearing, a comment shared by the ASAP Unit coordinator illustrated the ongoing confusion: "The teachers will still be making the decisions as to who graduates. But now we'll know the kids are competent."

As suggested by Madaus, stakes was a matter of perception. As some policy makers interpreted the ideals of the assessment policy as low-stakes, others relied on precedence and "technical expertise" to move the reform into the high-stakes arena. Amidst the cacophony of interpretations of the measurement-driven reform movement in Arizona, those who acclaimed higher standards and accountability overpowered the whispers of those seeking pedagogical change.

CONCLUSION

The incoherent messages sent by Arizona's policy makers to its audience—school teachers and administrators—are certain to effect myriad repercussions. Fuhrman (1993) attributed much of the failure of educational policy to improve education to inconsistency and lack of unified purpose. Arizona's plan to reform its schools is still held captive by the conflicting political forces and ideologies that influenced its creation. Although ASAP appeals to many because of its ambiguity, this same characteristic may undermine its capacity to effect any substantial change in educational practice.

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.

Baracy, J. M. (1992). *The political and social influence of the development of the Arizona Student Assessment Program.* Unpublished doctoral dissertation, Arizona State University.

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.

Finch, J. (1986). *Research and policy: The uses of qualitative methods in social and educational research.* London: The Falmer Press.

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

Fuhrman, S. H. (Ed.). (1993). *Designing coherent education policy: Improving the system.* San Francisco: Jossey-Bass.

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.

Hatch, A., & Freeman, E.B. (1988). Who's pushing whom? Stress and kindergarten. *Phi Delta Kappan, 70*(2), 145-147.

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

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.

Lindblom, C. E. (1980). *The policy-making process* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

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

Madaus, G. (1988). The influence of testing on curriculum. In L. Tanner (Ed.), *Critical issues in curriculum: 87th Yearbook of the NSSE, Part 1*. Chicago: University of Chicago Press.

Majchrzak, A. (1984). *Methods for policy research*. Beverly Hills, CA: Sage Publications.

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

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

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

Palumbo, D.J., & Calista, D.J., Eds. (1990). *Implementation and the policy process: Opening up the black box*. New York: Greenwood Press.

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

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

Rein, M. (1983). *From policy to practice*. Armonk, NY: M.E. Sharpe, Inc.

Resnick, L.B., & Resnick, D.P. (1989a). *Assessing the thinking curriculum: New tools for educational reform*. 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.

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

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

Strauss, A.L. (1987). *Qualitative analysis for social scientists*. New York: Cambridge University 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.

Weatherley, R., & Lipsky, M. (1978). Street-level bureaucrats and institutional innovation: implementing Special Education reform. *Harvard Educational Review*, *47*(2), 171-197.