DOCUMENT RESUME

ED 377 234 TM 022 495

AUTHOR Corbett, H. Dickson; Wilson, Bruce L.
TITLE Unintended and Unwelcome: The Local Impact of State Testing.
INSTITUTION Research for Better Schools, Inc., Philadelphia, Pa.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Apr 90
NOTE 41p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Educational Change; Educational Policy; Elementary Secondary Education; *Local Issues; *Minimum Competency Testing; Policy Formation; *Program Implementation; School Districts; Standardized Tests; *State Programs; *Testing Programs; Test Results; *Test Use
IDENTIFIERS High Stakes Tests; *Reform Efforts

ABSTRACT

This paper summarizes the results of a study of the local consequences of implementing statewide minimum competency tests. For American education to be the best in the world, the use of statewide and nationwide standardized testing as a primary policy tool for stimulating reform must be discontinued. Second, school district responses to such testing generally do not represent improvement and they certainly have not resembled reform. In the third place, this lack of a reform-like response may be interpreted as either a misuse of testing on the part of educators or a misuse of testing as a tool of reform on the part of policymakers. Fourth, testing is misused by policymakers because: (1) measures of student weaknesses are not adequate measures of system weaknesses; (2) uniform measures ignore important differences among districts; and (3) testing policies tend to engender conditions at the local level under which the reform intentions of the policy become unrealizable. Fifth, continued use of testing in educational change requires the recognition that policies must be established at the levels where action is to occur. Educators must do a better job selling education to the public and convincing the public that educational change is a long-term process, not a short-term improvement in test scores. (Contains 35 references.) (SLD)

TM

# UNINTENDED AND UNWELCOME: THE LOCAL IMPACT OF STATE TESTING



H. Dickson Corbett

Bruce L. Wilson

April, 1990

Research for Better Schools, Inc
444 North Third Street
Philadelphia, PA 19123

This paper is being presented as part of a symposium entitled, "The State Search for Indicators of Educational Performance" at the Annual Meeting of the American Educational Research Association, Boston, April 16, 1990. This paper is based upon Chapter seven of our forthcoming book, Testing, Reform and Rebellion being published by Ablex as part of their "Interpretive, Perspectives on Education and Policy Series" edited by George W. Noblit and William T. Pink.

The preparation of this paper was partly supported by funds from the U.S. Department of Education, Office of Educational Research and Improvement (OERI). The opinions expressed do not necessarily reflect the position of OERI, and no official endorsement should be inferred.

UNINTENDED AND UNWELCOME: THE LOCAL IMPACT OF STATE TESTING

This paper summarizes what we have have learned from a study of the local consequences of implementing statewide minimum competency tests (Corbett and Wilson, 1990). Unquestionably, such tests can force school districts to act on the basis of the results, particularly when the perceived level of stakes and the amount of pressure on districts to raise test performance are high. Unfortunately, the actions taken generally do not represent what educators themselves call "improvement" nor are the actions the product of planning processes that tend to generate high quality decisions or staff commitment to implementing changes. If high stakes-high pressure testing situations cannot encourage improvement, then it is unlikely that such an approach will be an effective stimulus of more broad-ranging reform; and yet that approach, embodied in "reform by comparison" thinking, has achieved a prominent place in state and national educational policymaking (e.g., Needham, 1989).

The paper's argument can be stated in five sentences. First, for American education to be the best in the world, as the participants in the Fall of 1989 presidential summit called for, we have to discontinue the use of statewide and nationwide standardized testing as a primary policy tool for stimulating reform action at the local level. Second, school district responses to such testing programs generally do not represent improvement and certainly do not resemble reform. Third,

this lack of a reform-like response may be interpreted as either a misuse of testing on the part of educators or a misuse of testing as a tool for reform on the part of policymakers; and we submit that the latter is the case. Fourth, testing is misused by policymakers in three ways: (1) measures of student weaknesses are not appropriate as guides for correcting system weaknesses; (2) uniform measures ignore important differences among school districts; and (3) testing policies tend to engender conditions at the local level under which the reform intentions of the policy become unrealizable. Fifth, there are three implications of this argument for the continued use of testing in reforming education: (1) policies requiring action must be established at the levels where action is expected; (2) educators must do a better job at selling education to the public, especially in terms of identifying what it is they should be held accountable for accomplishing; and (3) we all must realize that substantive educational improvement -- as opposed to educational change -- takes place over the long-term and, therefore, must concentrate on altering the aspects of reform policy that pressure educators to improve test scores in the short-term.

This paper will address each of the above five sentences in greater detail.

## The Call for Educational Reform

The idea of reform is that school systems need to do more than

simply become better at doing the jobs they currently are doing. School systems need to rethink their purposes, structures, and processes in order to create new jobs to do and to develop new "forms" of schooling that are more appropriate for enabling diverse student populations to function successfully in a knowledge-based society. The success of state reform initiatives, then, has to be assessed by the extent to which these efforts encourage, cajole, or force school districts to reconsider what their educational enterprise is all about, how they need to organize themselves, and what they have to know and be able to do in order to be successful.

For example, the present emphasis on redesigning, or "restructuring," schools recognizes that a comprehensive approach to reform involving professional educators, students, lay boards, parents, and citizens is much more likely to be effective in the long run than is occasional patchwork improvement of a program here or a procedure there. "Restructuring" means changing a school district's patterns of rules, roles, and relationships for the purpose of producing substantially different results. In other words, to do a significantly different and better job of educating students, school systems must alter (1) existing, shared understandings that direct the operation of the local educational enterprise, (2) the regular ways in which students, teachers, and administrators carry out their jobs, and (3) the ways in which these people are accustomed to responding to one another. The concern in "true" restructuring efforts must not be with

just how particular policies, programs, or practices should be implemented but also with how the school district itself can function to reinforce, rather than contradict, those changes. That is, the structure of schooling must reinforce the process of schooling which in turn improves the effectiveness of schooling.

Many educators seem to agree that such reform is necessary. Witness the proliferation in the late 1980s of broad-ranging and dramatic attempts to design forms of education that break with the past: the National Education Association's Learning Laboratories Initiatives (see NEA, 1990); the American Federation of Teachers' Professional Development Schools (Viadero, 1990); the Coalition of Essential Schools program led by Ted Sizer (see Sizer, 1989); the Accelerated Schools Program initiated by Henry Levin (Levin, 1987); the work of the Center for Leadership in School Reform under Philip Schlechty (Schlechty, Ingwerson, and Brooks, 1988); James Comer's partnership program of teachers (Comer, 1988); not to mention the national attention-drawing programs in Dade County (Florida), Rochester (New York), Hammond (Indiana), New Orleans (Louisiana), Cincinnati (Ohio), Jefferson County (Kentucky), East Harlem (New York), and Cerritos (California) (see David, 1989).

Disagreement arises over the means of achieving reform goals, however. One such means is the use of statewide minimum competency tests as a mechanism to increase accountability. To what extent do statewide minimum competency testing policies promote the kind of

serious considerations necessary to stimulate educational reform? Our answer is that these policies do not promote serious reexamination and, indeed, engender responses that are contradictory to reform. There is no question that policymakers can control education through such policies; there is serious doubt that they can reform education positively through such policies.

It may be that the reader regards the imposition of purpose from the outside as a necessary step in the reform of American education and the removal of contextual differences in indicators of satisfaction with local schools as positive developments. Even if that is so, the use of current statewide testing programs as policy tools to accomplish that vision is not the means to do so.

The paradox is that the success of the current reform movement is likely to be determined by how well the issue of assessing results is handled by educators. Traditional measures and existing assessment programs (such as many of the statewide tests currently in place) were created under traditional assumptions about the purpose of schooling and how schooling occurs. To the extent that these devices guide a system toward improvement, they are likely to guide the system to do better at what it is already doing. The more legitimate purpose of reform is to enable schools do a job they have never done before. Contradictions between current assessment strategies and this purpose are major obstacles to reform.

Values and beliefs about the purposes of education lead to rules

that govern the behavior of educators which in turn lead to the forms that schooling takes (structure) and how schooling is conducted (process). Educational reform, by nature, calls into question those values and beliefs about education that have given rise to the current forms schooling takes. (Schooling and education are not used interchangeably here following Willard Waller's observation that it is possible to love education and hate school [Cohen, 1989].) To promote educational reform then requires one to promote a different set of values and beliefs about education.

Test results do not lend themselves to encouraging an alteration in structure or process, nor do they lend themselves to stimulating a reexamination of educational purposes that might lead to a different formulation of structure and process. Indeed, existing statewide tests reinforce a number of traditional values and beliefs about education -- namely, that there is a body of content students must master and a set of skills students must demonstrate by a particular age, that this content and these skills can be reflected in student responses to paper and pencil tests, and that student failure to respond successfully on tests is the school's responsibility to correct. These values and beliefs have led educators, and the communities within which they work, to construct a form of schooling that has certain distinguishing characteristics (despite the presence of variations from school district to school district). These characteristics, often noted in critiques of schooling (see Jackson, 1968; Dreeben, 1968; Schlechty,

1976), include adherence to age-grade distinctions, mostly passive involvement of students in the educational process, and clear delineation of specific courses defined by content areas.

There will be no room for reforming schooling until the purposes of education are rethought. In other words, reform begins with purpose, not outcomes. It is only after purpose has been clarified well that it makes sense to discuss what the results of enacting this purpose should be. Even if schools got much better about serving the educational purposes represented in current statewide testing programs, the needs of society would not be served. Not only should schools do a better job at what they currently do, but they need to begin doing jobs they have never done before. Testing, done in the best possible way, is only a tool for the former; the latter is instigated by serious examination of the purpose of education. Thus, testing has a purpose in reform, but it is not as stimulant; rather it is an informational tool, one among many, about how to adjust schooling to enable students to learn better.

In fact, there is a danger in school districts' relying on standardized tests as one set of outcomes for which they hold themselves accountable, if the districts are serious about reform. An almost inevitable lack of congruence exists between the reformulated purposes of education such districts pursue and the more traditional purposes of education such statewide tests target. Districts will find that they are accountable for student outcomes that are not necessarily

the logical consequences of the purposes they now seek. Given that goals and outcomes can be equally compelling forces for directing organizational behavior (Mintzberg, 1983), the public pressure that has been shown to accompany the improvement of high-stakes test results will probably supply enough impetus to resolve the contradiction in favor of the purposes embedded in the tests. In this way, testing actually becomes a tool that blocks reform efforts.

## School District Responses and Reform

Is there any evidence that school districts, as a response to participating in statewide minimum competency testing programs, have rethought the purpose, structure, and process of schooling in their community? The answer, based on the evidence based on our research is no (Corbett and Wilson, 1990). The empirical data indicate that school districts' responses to the tests are, at best, conservative ones and actually represent a form of rebellion against reform. The responses are frequently crisis-oriented, formulated to appease various stakeholders in the educational system, rather than a system-oriented one that is intended to revamp the district.

Our survey findings (see Chapter Four of Corbett and Wilson, 1990) found that, in general, statewide tests which are tied to student performance consequences (e.g., graduation) have a greater effect on a school district's organization (particularly the use of test results as accountability benchmarks), technology (in terms of strategies used to

address the test content and changes in curriculum and instruction), and culture (as defined by characteristics of teachers' and students' worklives). Interviews probed further into the contextual conditions that produced these effects and the findings indicated that as the perceived level of the stakes associated with a test increased and the pressure on a district to improve its performance mounted, the effects of implementing the testing program were more of the sort associated with raising scores rather than improving learning. Indeed, the professional educators themselves noted that there was a point where strategies to raise scores diverged from strategies to improve learning. Thus, under conditions of high stakes and high pressure the effects of statewide testing were not synonymous with greater educational effectiveness.

Coping with the pressure to attain satisfactory results on high-stakes tests caused educators to develop almost a "crisis mentality" in their approach, in that they jumped quickly into "solutions" to address a specific issue. They narrowed the range of instructional strategies from which they selected means to instruct their students; they narrowed the content of the material they chose to present to students; and they narrowed the range of course offerings available to students -- to the point that for those students who had difficulty passing the tests one could almost say there was an informal remediation "track."

To many educational observers and policy makers, this narrowing is

not a negative development (see Murphy, 1990 for a summary). For them, it represents a focusing of the instructional program, ridding the curriculum of the distractors that have prevented schools from doing the job of providing all students with essential learning skills. Indeed, the implementation of these testing programs made "back to basics" a reality for many school districts, even for a good number of educators who were not advocates of that philosophy.

Although this ability of a statewide testing program to control local activity may be praiseworthy in the minds of some educational critics, the activity the program stimulated was not reform. Responding to testing did not encourage educators to reconsider the purposes of schooling; their purpose quickly became to raise scores and lower the pressure directed toward them. Responding to testing did not encourage educators to restructure their districts; they redirected time, money, and effort so that some parts of their systems could more expeditiously address the test score crisis while leaving the parts unaffected by testing or producing "good" scores unscathed. Responding to testing did not encourage educators to rethink how they should teach or how they should administer schools; once again, they addressed process only in the parts of their system that felt the direct impacts of testing. Responding to testing did not encourage educators even to reaffirm existing purposes, structures, and processes as efficacious; they rarely, if at all, seriously considered the alternatives.

Instead, educators relied on instructional and organizational

"habits" that had been present in their educational systems for a long time -- e.g., drill and review in classrooms, pull-out programs for remedial instruction, assigning additional duties to existing positions, etc. -- even though some of the habits, particularly those related to instruction were ones that many educators believed did not represent state-of-the-art practice. Thus, the majority of effects we observed represented an eschewing of systematic analysis of alternative educational purposes, structures, and processes and a reinforcement of educational practice that had been present in American education for years.

The data show, however, that even if one subscribes to the view that conforming to this model of education is the right thing to do, educators themselves suggested that they were doing the "right" thing for the wrong reasons. And there was a considerable element of the educational community we talked to who believed the testing program had put them in the position of doing the wrong things for the wrong reasons. Thus, a peculiar form of rebellion seemed to be present in the local responses wherein school districts did not simply reject the goals and means of testing in favor of what they considered more appropriate goals and means. Rather, in the face of high stakes and high pressure to increase results, they focused more squarely on the results as the ultimate outcome of their activity and more exclusively on those instructional means that had the quickest payoff in terms of improving results: repetition, review, and remediation.

We have hesitated to label any of these practices as "teaching the test" -- or even its more semantically palatable relative, "teaching to the objectives of the test." Those terms are so value-laden that they are often considered to be euphemisms for out and out cheating. Koretz (1988:15) makes a much more useful distinction and categorizes such practices as those that (1) "inflate" test scores while "degrading" the curriculum, (2) inflate test scores while leaving the curriculum unimproved but unharmed, and (3) inflate test scores while advancing the curriculum. Most of the districts we studied probably engaged in all three at worst. That is, no district could be construed as an unequivocable "cheater." As the perceived level of the stakes and the pressure to perform increased, the mix of the three types of practices changed, with a greater proportion of the first two and less of the third. Regardless of the intentions of policymakers in initiating the testing programs or of the "goodness" of educators' reponses, empirically we found that school districts responded to statewide testing in ways that did little to reform the way they practiced education.

## Educators or Policymakers as the Source of "Blame" for the Lack of Reform?

From a theoretical standpoint, we are interested in whether these identified effects are the products of "educational misuse" of the tests or the inevitable consequences of "policy misuse" of testing.

Educational misuse means that the professional educators are either allowing test performance to influence decisions that the tests were not intended to influence or engaging in practices that can "inflate" test performance but "degrade" curriculum and instruction, to use Koretz' (1988) terms. To say that educators have misused the tests is to say the problem resides in educators' behavior. While we will argue that the preponderance of the effects we saw were not the result of educational misuse, we do not deny that such misuse did occur. For example, after we presented several of the findings from this study to an audience of Maryland teachers, one of those attending relayed a story to us about a principal who reassigned teachers to unattractive positions based on student performance on the statewide test in the teachers' respective subject areas. The teacher telling the story was one of those who was reassigned. If such "implementation problems," as they are often referred to by policymakers, were widespread, then appropriate solutions to the emergence of "negative" consequences would be to have state education agencies provide additional information and assistance to local school districts concerning proper use of the scores or create more rigid state monitoring of test administration and interpretation, all the while leaving the tests unscathed.

Policy misuse means that educational policy makers are using an inappropriate lever for instigating reform at the local level. In this scenario the blame shifts. The problem is not that educators subvert policy intentions but that the policy tools themselves contradict the

intentions. What we found is that the high-stakes, high-pressure environment created by statewide testing programs encourages rebellion against the very reform goals the policies sought to attain. The following section provides a critique of testing that substantiates our contention that existing statewide standardized tests are inappropriate for stimulating reform.

## Reasons for Policy Misuse of Testing

Debate surrounding the role of statewide testing in reform must try consciously to avoid praising or damning statewide minimum competency tests as significant promoters of reform just because of their observed effects on local school districts. The question is whether this conservative (and rebellious) response is poor implementation, i.e., the product of educators' "misuse" of the tests, or whether there is something intrinsic to statewide minimum competency testing policies that renders them inappropriate as tools for reform. This section argues that statewide testing of student learning outcomes actually represents "policy misuse" because such policies are inherently poor stimulants of local reexamination of purpose, structure, and process for three reasons:

- Outcome measures stated in terms of student learning do not provide direction as to what school systems should do differently to produce different results.
- Testing programs, both in terms of results and in the implications for action, ignore variations in district contexts that may affect the importance of the results and the appropriateness of certain responses from community to

community.

- Statewide testing policies tend to foster conditions antithetical to actual reform.

Each of these reasons is discussed in more detail below.

Test Results As Inappropriate Guides for Action

The growing evidence is that state-mandated minimum competency tests can control activity at the local level. Apart from our research, others have found this to be the case as well. For example, Dorr-Breeme and Herman (1986) note that local districts are emphasizing locally-developed curriculum and instruction objectives less and less and instead are beginning to rely on those embedded in state tests; and Tyson-Bersten (1988) reports that textbook selection in states that have statewide adoptions are beginning to be made primarily on the basis of whether the books fit the test.

While testing programs can control activity at the local level, the kind of statewide tests with which we are concerned are not useful for guiding reform activities. Information about student learning outcomes contained in minimum competency test results only provide information about what it is students do not know or do not know how to do. This information tells educators nothing about how the school system should be organized and operated differently to alter those outcomes, should they perceive that the results are less than desirable. To the extent that test results guide action at all it is in the direction of working harder at doing what schools have already

been doing. To the extent that what schools have already been doing contains a considerable number of "bad" habits in their preparation of students to live effectively in modern society, then school district staff members will essentially latch onto their bad habits more intensively. The fault is not the educators'; they have no guidance from the indicators imposed on them by local and state policymakers to do otherwise.

Suppose, for example, students in third grade are having difficulty finding the least common denominator in adding or subtracting fractions. A "logical" conclusion would be that third grade teachers need to concentrate on that objective more. But, what does "concentrate" mean for the adults in the school system in terms of rethinking educational purpose, process, or structure? With respect to purpose, do the professional educators decide that this skill is superordinate to other skills that eight year olds should develop, or do they decide that spending more time on this particular skill would interfere with other, perhaps more important, priorities? Do they decide that the problem is one of process and engage in the search for different instructional techniques that would enable staff to teach math skills more effectively without having to allocate more time to th particular skill in question? Or do they venture into structural solutions and decide that self-contained, homogenous classrooms are incongruent with effective math instruction? Perhaps in the absence of any guidance they will adopt a "plumber's friend" style of reform and

do all of the above. The test score information offers no basis for such decisionmaking.

Certain test results may be phrased in terms of student performance; but indicators of the outcomes of student performance -- as opposed to indicators of the quality of the performance itself -- provide little guidance as to what it is about teacher and administrator behavior that has to be changed in order to improve student performance. If, as another example, a school district's staff members discover that 35 percent of its students have failed a state minimum competency test in reading, where do they turn for remedies? The test results (including detailed analyses of test objectives) do not tell them whether students need more reading instruction, different reading instruction, better reading teachers, increased opportunities to develop higher order thinking skills, or an improved classroom learning environment, to name just a few of the possible implications of poor reading scores. Student outcome measures, by themselves, are simply not useful for driving reform.

In Schlechty's (1990) view, student outcomes are the products of quality but do not measure quality themselves. Quality measures attend to the actual work that students, teachers and administrators perform. Thus, while a district will clearly have differences in student outcomes in mind when it undertakes its reform effort, it also will focus on a variety of intermediate steps related to student and staff performance, the attainment of which are assumed to lead to improved

student learning. Such results may be the extent to which students complete classroom and homework assignments, the amount of time students actually engage in school work, the development of a common language of instruction among all staff members, knowledge about and agreement with a shared purpose concerning the district's work and/or the quality of the work that staff members design for students to do.

For example, a principal in a Wisconsin high school explained to us during a conference that he and his staff believed that students' failure to complete classroom and home assignments was preventing them from learning as well as they should. They decided that one way to improve student learning, then, was to insure that all students completed all assignments and added an extra period at the end of the school day for every student who had not finished assignments as a result of this conclusion. Students who could do the work, but previously had not, quickly began to finish tasks on time; students who could not do the work were identified, given redesigned assignments, and/or provided special instruction. By altering the quality of student performance the faculty was able to alter a widely used indicator of student learning without concentrating on strategies that would inflate the test scores on that specific test. The result was that student achievement test results began to improve, but not because the school directed itself to the results of specific student assessments but because the school assessed an aspect of its operation and found it lacking.

Of course, the rejoinder is that information on student weaknesses should encourage educators to look for the right actions to take and that such information may be a necessary prelude to deciding there is a need to act. However, a specifically aimed kick in the pants will lead most school districts to address that problem specifically; and school districts that teach reading poorly will simply tend to teach reading poorly more intensively without focusing on more generic questions of reform. If most school districts knew the right actions to take, they would take them -- in the absence of pressure to behave in a contrary manner.

For the most part, educators know what many of the right actions to take are in terms of the process of changing. There is a history of school improvement research in this country that illustrates the processes that have to take place for change to be positive and successfully implemented (see Fullan, 1982; Huberman and Miles, 1984; and Lieberman, 1986). Moreover, there is considerable research on what school districts can do to improve instruction specifically, e.g., attending to differences in student and teacher learning styles, encouraging the classroom use of cooperative learning, reinforcing effective instructional techniques (normally associated with Madeline Hunter programs), and using peer coaching strategies.

If policymakers want to encourage school districts to take these "correct" actions, then perhaps they should use different comparisons than those provided by student test performance to force action. They

should make comparisons, for example, between school districts in terms of how much time administrators spend on instructional matters. They should make comparisons between school districts on the frequency and intensity of staff development opportunities for teachers. They should make comparisons on the amount of time that reading teachers in one school district spend talking with regular teachers versus the amount of time that reading teachers in another district spend talking to regular teachers. If one's retort to this line of argument is that it is not practical for states to gather information on such topics, then essentially one is saying that it is not practical for statewide policy makers to encourage reform, meaningful reform, at the local level through public comparisons.

If policymakers are serious about encouraging reform, they should establish comparisons that concern the behavior of educators and students. We argued earlier in the chapter that new patterns of rules, roles, and relationships are needed to produce different results; if so, then assessments should provide considerable direct information on what it is about those new patterns that is effective or ineffective. The point is that diagnosing student weaknesses is not the same as diagnosing system weaknesses; and without system diagnosis, little guidance is available as to what it is about existing purposes, structures, and processes that need changing.

Good strides have been made with respect to better assessments of the quality of student performance (Wiggins, 1990). For example,

Connecticut is beginning a performance-based assessment program in math and science. While several states have started similar experiments (e.g., California and New York), Connecticut's seems to be the first to move into large-scale testing of this sort. The program "will measure student performance on a series of tasks that may take as long as a semester to complete... Students will be asked to work individually and in groups to frame problems, collect data, and analyze and report their results" (Rothman, 1989:1,21). These means will do a better job at telling educators what students need to do differently.

However, similar developments have not been made with respect to assessments of what educators need to do differently. Even with current proposals for new teacher assessments in place (see Bradley, 1989 for a summary), there remains a lot of work that has to be done before "system assessments" will be available. At a minimum, a system assessment must attend to what administrators need to know and be able to do in order to support teachers' ability to obtain and use the knowledge and skills necessary to encourage students to behave in ways that lead to learning. That is, for an assessment to be of use in governing action, it must inform the system about relationships among elements of the district rather than just particular characteristics of certain elements in isolation from the others.

There are a growing cadre of testing critics who argue that what these tests measure is not what should be tested. Thus, even if such outcomes were capable of guiding action, they would be guiding action

in inappropriate ways. While these arguments have considerable merit, they still do not get around the problem of trying to direct district behavior through student outcome measurements. Such measurements cannot be very helpful to a school system without the presence of other indicators of the quality of the performance of the system.

As a final note, if one accepts that poor test scores indicate a need for reform action, then one must accept that "good" test scores are a reason to maintain the status quo. This "Don't fix what ain't broke" philosophy may mask considerable room for improvement by enabling system defenders to justify the appropriateness of current practices, even if the major responsibility for the high scores lies in the laps of an advantaged student population and not with the excellence of the professionals, programs, or practices.

School Context Variations

A second reason for why state testing policies are ineffective stimulants of local reform is their inability to accomodate local context differences. Mandated statewide testing policies have a resilient insensitivity to the context-bound nature of school improvement, primarily because they treat the results as universally applicable and suppress exceptions to the need for implementing the program. According to McDonnell and Elmore (1987:140),

> Mandates assume (a) that the required action is something all individuals and agencies should be expected to do, regardless of their differing capacities, and (b) that the required actions would not occur with the frequency or consistency specified by the

> policy, in the absence of explicit prescription. Rules, in other words, are introduced to create uniformity of behavior or, at least, to reduce variations in behavior to some tolerable level.

School districts serving a disproportionately large group of disadvantaged students are compared with those districts that have predominantly advantaged students; districts with limited resources must find ways to remediate its students who fail the tests even though districts with more resources may have even fewer students to instruct; and districts with local testing programs aligned with their local curriculum must implement the state program just as those districts with no local tests must. This press for uniformity is understandable in political terms, but it leaves little room for school districts to determine which results they wish to address (or whether they even need to do so) or to adopt a timetable for reform that best fits the exigencies imposed by local contextual conditions.

A school district is not a school district is not a school district. Certainly there are similarities, but each system has a different set of conditions that it faces. These differences make a difference in terms of what needs to be done, how to do it, and what the outcomes of the doing will be. In fact, two of the major themes in the literature on school improvement are (1) that some changes work some times in some places and (2) that being sensitive to the local context in which implementation is to occur will enable leaders to improve their chances of making changes that are effective in that setting and that last (Berman, 1981; Corbett, Dawson, and Firestone,

1984).

Context, then, refers to the time and the place in which a policy is to be implemented. "Time" is important because there is a "right" time for change at the local level, or at least an expected time for it. For example, teachers often anticipate, although not necessarily welcome, a period of change with the advent of a new principal, superintendent, or school board. Or, occasionally, educators who begin to share a sense that their work has become stagnant also begin to share an expectation that change should be imminent. Of course, the time for change is not right in all districts at the same time. The use of statewide testing programs as leverage to force local activity is oblivious to such differences. Districts that would welcome an external stimulus to "jumpstart" an improvement effort, districts that already have embarked on reforms of their own and would be resentful of external attempts to disrupt their activity, and districts that have a considerable number of intractable problems to address before a concentrated focus on instruction can be mounted are all treated equally in the eyes of policymakers.

There is also a certain amount of time needed for changes to occur and take hold. It may take as long as three to five years for significant educational reform to be planned, implemented, revised, and finally incorporated into the system -- depending on the situation of the individual districts (Fullan, 1985). On the other hand, yearly testing cycles create a sense in the public's mind that improvement

will be reflected in steadily increasing test scores. If a district engaging in longer-term reform activity is not fortunate enough to enjoy a serendipitous increase in test scores, the reform effort is not likely to be buffered long enough to have a chance to be institutionalized. There is no guarantee, in fact, that the disruption that accompanies wholesale changes in an organization will not actually cause test scores -- and other more subjective outcome measures -- to decline for a while (Eastwood, 1990). Thus, yearly high-stakes testing may actually mediate against the engagement in practices that ultimately will be the most effective in improving the system.

The "place" is also an important aspect of school context and includes a district's organization (structure and process), culture (definitions of what is and what ought to be), politics (distribution of power and decisionmaking authority), and economics (availability and allocation of resources). As with time, districts vary on these characteristics and, moreover, schools within districts can vary on these characteristics. Statewide testing policies such as the ones we studied impose a uniform need to respond on systems that have widely disparate capabilities to respond. Some are able internally to limit the external attempt to control their activity by encapsulating the response and continuing life as usual elsewhere in the system; others are unable to resist external demands and thus are continually tossed about on successive waves of initiatives. McDonnell and McLaughlin (1981) label the former local systems as "independent actors" and the

latter as "junior partners," designations that reflect differences in local contexts and subsequent differences in the capability to act across various settings.

Of the ways that "places" can vary, perhaps the most important to consider is culture. Culture is "shared definitions of what is and what ought to be, symbolized in act and artifact" (Wilson, 1971:91). These definitions embody statements of purpose about what educational activity in a particular subject department, school, or district seeks to accomplish. It follows that if definitions of purpose differ from setting to setting, then the indicators educators and citizens use to determine the degree of their satisfaction with their pursuit of purpose will also vary from setting to setting. Districts with many college-bound graduates will likely keep a close watch on SAT scores; districts with many drop-outs will likely scrutinize students' reasons for leaving school and the overall dropout rate to determine how well they are doing. The use of statewide tests as a control mechanism tends to encourage the use of the results as benchmarks of performance (Corbett and Wilson, 1990). The results then become an all-purpose indicator, irrespective of the possibility that the indicator is inappropriate for locally-defined purposes. The consequence is that a potentially disruptive incongruence emerges between purpose and results. Thus, statewide testing programs, in addition to being inappropriate as guides for reform, are too "blunt" to encourage widespread purposeful activity. Instead, for many

districts, the initiative will set in motion activities that distract educators from their purposes.

Creating Conditions That Engender Responses Contrary to Policy Intentions

The third way in which policymakers misuse testing as a part of educational reform policy is that they tend to create conditions that encourage local responses that are contrary to the reform intentions of the policy. A policy must be considered as both the intentions for local activity set forth in the policy and the conditions the policy engenders under which districts are expected to act. The issue really is one of control; and in attempting to control educators' responses, policymakers tend to force them to respond in ways that are not in line with reform.

To say that the effects are the product of the testing policies is to say much more than the tests may not actually measure learning appropriately. Testing and measurement experts have a very active debate in progress among themselves regarding these technical issues (e.g., Tittle, Kelly-Benjamin, and Sacks, forthcoming; and McLean, forthcoming). But testing policy includes not only the instrument and its characteristics but also, and in our minds more importantly, the testing situations it engenders at the local level. This is an important point because it recognizes that policy can establish relationships between policy intentions and target consequences that

shape the range of possible consequences. Merton's (1968:106-107)) comment concerning the relationships among elements of a social structure is salient:

> The range of variation in the items which can fulfill designated functions in a social structure is not unlimited ...The interdependence of the elements of a social structure limits the effective possibilities of change or functional alternatives...Failure to recognize the relevance of interdependence and attendant structural restraints leads to utopian thought in which it is tacitly assumed that certain elements of a social system can be eliminated without affecting the rest of that system.

Implementing a testing policy that affects student progress in school and is amenable to public comparison of test results sets up a probable sequence of events leading to the development of a local perception that the stakes are high and increased pressure on schools to perform. In turn, these conditions engender the kind of response we have labeled as rebellion (Corbett and Wilson, 1990). This probable sequence of events is a direct consequence of establishing testing policy in a certain way. When the probable occurrence of local phenomena that can affect the realization of policy objectives is heightened by the policy itself, then the local consequences of enacting the policy should be considered a "policymaking problem" and not an "implementation problem." To continue to focus on educators as the solution to problems generated by the testing policy is akin to blaming students for failing to learn from bad teaching.

This situation of creating conditions that make it impossible for the intended response to occur is perhaps best illustrated through a

personal example. One of us lives in a two-career, three-kid house; and for life to run smoothly in the mornings, the parents established the policy that the kids will dress themselves in the morning in clothes that are appropriate for school. "Appropriate" is defined broadly: anything that was not slept in or selected from the dirty clothes basket. To make things even easier, the parents usually set these clothes out the evening before, so all a child has to do is get up and get dressed. However, when the husband walks the hall in the mornings, gazing into bedrooms, he often sees signs of noncompliance. With the six year-old, he begins to encourage compliance with a few gentle reminders. As time passes and the need to get everyone in the car becomes more urgent, he becomes more forceful -- and resorts to the "implied threat", sternly stated: "You better get dressed NOW!" Unfortunately, this statement usually initiates "the whine." "The whine" begins with a low whimper, and for a parent, it is more grating than a finger nail on a blackboard. The whine takes enough kid concentration so as to preclude any immediate compliance and, thus, is quickly followed by the "direct threat": "Get dressed or you will .....(to be filled in by a promise to deprive the child of what she most wants)." This more intensive attempt at controlling behavior stimulates the "all-out wail," and the child becomes a shuddering mass of tears, incapable of doing anything. So either the parents dress her as she cries (contrary to their policy's intentions) or she eventually throws on her favorite play clothes that have dodged the washing

machine for weeks (again contrary to the policy of appropriate dress for school).

Educational policymakers behave similarly -- by holding up certain tests as the primary public indicators of the quality of schooling and mixing in intense pressure on social districts to improve test scores immediately. When one combines those conditions with yearly testing cycles, one creates circumstances under which educators will improve scores without improving learning; one creates situations in which educators would tend to do all the things that are contrary to what the vast knowledge we have about school improvement says is necessary to significantly improve schools. That is, educators will try to make changes quickly, look for an immediate impact, and limit the involvement of others in decisionmaking. All three of those strategies lead directly to short-term change without lasting long-term improvement.

Sykes and Elmore (1989) offer additional insights into how testing policies generate conditions that hinder, rather than help, reform:

- Attention to indicators of organizational performance focus on the indicators themselves and not the underlying purposes of education; that is, "test results become the result, not learning."
- What is measured draws attention from that which is not measured, thereby putting a host of other, important educational purposes in the backseat.
- Testing encourages a standardized pedagogy for use with a diverse student population.
- A focus on test results ignores alternative means of attaining the broader learning results and, consequently, offers few

rewards for innovation, risk taking, or entrepreneurship.

- An accountability-driven system distorts the motivational climate for teaching and learning.

The problem, thus, is that many policies with stated intentions of long-term reform use tools that force specifically-targeted, short-term responses. The blame for lack of reform should be placed at the feet of the policymakers and not at the feet of the policy implementers. To attack educators for failing to respond heroically to a policy initiative in which the conditions for implementation contradict the intentions is to use a logic that blames students for failing to learn in the presence of poor instruction.

## Implications for Testing and Reform

Essentially we see three implications of the above argument for the making of educational reform policies and the place of statewide standardized testing in those policies. The first concerns the need for the locus of decisionmaking to be the school district and is best illustrated by the end of the story about one of our daughters and the morning crisis. After trying futilely for months to get her to comply with the parents' desire for her to dress herself appropriately so the family could get out of the house in the morning on time, it occurred to the father to ask the daughter: "Why?" She calmly explained that the reason was that sometimes the parents would pick out shirts that would have a ribbon on the front, near the collar. When she would see that ribbon, she would remember that the ribbon would scratch her neck,

and when she thought of the ribbon scratching her neck all day in school, she did not want to put on the shirt. But when she did not put on the shirt, the father would start yelling at her, and she would whine and start crying; then, she would be crying so much that she could not put on the shirt at all. She concluded by saying, "Dad, why don't we pick out a shirt for me to wear?" Notice that she did not say, "let me choose." She was willing to allow her father to establish a range of acceptable dress from which a joint choice could be made. All she wanted was to assist the decision. Essentially she was saying, "Dad, you're creating conditions which make it impossible for me to do what it is you want me to; if you want me to do something, let me help decide what I should do."

That somewhat humbling lesson would seem to apply to educational reform policy as well. Policies should be set that involve the levels of the system where action is expected in the decisions. If we want local educators to engage in educational reform, then they must be highly involved in making the decisions about what kinds of reform that should and will take place. Educators should make the decisions about what they should be held accountable for, and then they should also be held accountable for going out and taking the actions that will enable them to improve on that indicator of accountability that they have selected. Decisions about what to be accountable for and how to meet that accountability should be the province of the local school system.

The state's role would be to establish the expectation that all

school systems would engage in this decisionmaking process responsibly. That is, educators would select or invent the particular instruments they want to use in their school districts to assess their progress toward meeting the accountabilities. The state policy simply would be that school districts must have some sort of means of showing what, and how well, students are learning. By enabling multiple tests to be used across the state, the policy would weaken the viability of those tests as means of comparison across school districts -- and across states.

The second implication is that educators must do a better job of selling to the public what they decide they should be held accountable for. For example, a superintendent taking a new job in one school district told the Board hiring him that it should establish accountabilities for him to meet and, then, suggested the accountabilities the Board should use. He essentially established his own accountability system. This particular school superintendent was interested in establishing site-based management as the mechanism through which schools would make decisions about how they would improve. He, therefore, established in his contract that his progress as a school superintendent would be measured by the extent of school progress toward implementing site-based management. If the Board made decisions that hindered the schools' progress, then the superintendent could argue that the Board was making it impossible for him to accomplish that for which they were holding him accountable. In essence, he convinced the Board what he should be held accountable for

and then held the Board accountable for not putting obstacles in the way of these agreed-upon goals.

There are people that advocate weakening the impact of a particular set of test scores on school districts by holding school districts accountable for a variety of indicators, thereby reducing the importance of any one particular indicator. We disagree with this. If everything is important, then nothing is important. If educators select a whole host of indicators as priorities, then no activity will be a priority in practice. Resources are simply too scarce to enable educators to focus on a wide variety of indicators of school quality. They must decide what is most important, select an indicator or two as the ones they will work on for a period of time, and convincingly communicate those priorities to the public.

Third, and finally, there has to be a focus on reducing the pressure on school districts to improve test scores in the short-term. This may happen in a number of ways. For one, yearly test administrations might be switched to every three, four, or five years. When yearly improvement on test scores is expected, then the change process is unwisely confined to a year's cycle. Significant change can not take place in a year, and the early phases of the process actually may be associated with unimproved or decreased scores as educators focus their attention on initial issues that may be removed from classroom practice. With a longer interval between test administrations, it is more likely that the tests will be used as

guages of progress rather than the sole indicators of success.

## Conclusion

By using statewide testing programs that encourage the perception of high stakes and the presence of high pressure to perform well on the test at the local level, policymakers create a situation where tests that have no capability of directing real reform will, nevertheless, stimulate local activity anyway. The critical point to remember is that the testing policy is not just the test itself. The policy includes the consequences for students and educators attached to particular performance levels on the test (both intended and unintended), the level of emphasis the state education agency and the legislature give to the test results (which signals how important the test is politically), and the various administrative demands associated with the testing program (e.g., the testing cycle, reporting procedures, funding for remediation, etc.). The mix of these characteristics helps to create the local conditions of stakes and pressure under which school districts will have to act. The policy cannot be separated from the probable sequence of events it will engender. If a district or two "cheats" out of a population of districts that are for the most part forthrightly trying to improve, then the one or two outliers represent an "implementation problem." But when the modal response to statewide testing by professional educators is typified by practices that even the educators acknowledge

are unwelcome and counterproductive to improving learning over the long term, then the issue is a "policymaking problem." To maintain the policy in such a situation and label the response as educational "misuse" is irresponsible.

The real educational question is how to accumulate evidence of how well schools are doing in such a way as to encourage, reward, and avoid obstructing honest and reasonable attempts at improving schooling. It is our perspective that testing programs which stimulate perceptions of high stakes and intense pressure to improve test performance have no part in policy initiatives that have educational reform as their purpose. The combination of high stakes and pressure create conditions where unintended and unwelcomed responses become the norm.

In interviewing educators in one state, we came across a response that captures the essence of our argument: "Statewide tests have taken an educational tool for improvement and turned it into a political weapon." Educational tools can help guide reform; political weapons can only force change. And there is a substantial difference between reforming schools and simply changing them.

## REFERENCES


Berman, P. E. (1981). Educational change: An implementation paradigm. In R. Lehming and M. Kane (Eds.), Improving schools: Using what we know. Beverly Hills, CA: Sage.

Cohen, D. (1989). Willard Waller, on hating school and loving education. In D. Willower and W. Boyd (Eds.), Willard Waller on education and schools. Berkeley, CA: McCutchan.

Comer, J. P. (1988). School power: Implications of an intervention project. New York: Free Press.

Corbett, H.D., & Wilson, B.L. (1990). Testing, reform and rebellion. Norwood, NJ: Ablex.

Corbett, H. D., Dawson, J. A., & Firestone, W. A. (1984). School context and school change. New York: Teachers College Press.

David, J. (1989). Restructuring in progress: Lessons from pioneering districts. Washington, DC: National Governor's Association.

Dorr-Breeme, D. & Herman, J. (1986). Assessing student achievement: A profile of classroom practices. Los Angeles: UCLA Center for the Study of Evaluation.

Dreeben, R. (1968). On what is learned in school. Reading, MA: Addison-Wesley.

Eastwood, K. W. (1990). Increasing the chances of successful school change: A model for lasting change. Paper presented at the Annual Meeting of the American Educational Research Association, Boston.

Fullan, M. (1982). The meaning of educational change. New York: Teachers College Press.

Fullan, M. (1985). Change processes and strategies at the local level. Elementary School Journal, 85(3), 391-422.

Huberman, M., & Miles, M.B. (1984). Innovation up close: How school improvement works. New York: Plenum.

Jackson, P. W. (1968). Life in classrooms. New York: Holt, Rinehart, & Winston.

Koretz, D. (1988). Arriving in Lake Wobegon: Are standardized tests exaggerating achievement and distorting instruction? American Educator, 12(2), 8-15, 46-52.

Levin, H. (1987). Accelerating schools for disadvantaged students. Educational Leadership, 44(6), 19-21.

Lieberman, A. (Ed.) (1986). Rethinking school improvement: Research, craft and concept. New York: Teachers College Press.

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

McDonnell, L., & McLaughlin, M. W. (1981). The state role in education: Independent actor or junior partner? Paper presented at the Annual Meeting of the American Political Science Association, New York.

McLean, L. (forthcoming). Pedagogical relevance in large-scale assessment. In R. Stake (Ed.), Effects of changes in assessment policy. Greenwich, CT: JAI Press.

Merton, R. (1968). Social theory and social structure. New York: Free Press.

Mintzberg, H. (1983). Structure in fives: Designing effective organizations. Englewood Cliffs, NJ: Prentice-Hall.

Murphy, J. (1990). The educational reform movement of the 1980s: A Comprehensive analysis. In J. Murphy (Ed.), The reform of American public education in the 1980s: Perspectives and cases. Berkeley, CA: McCutchan.

National Education Association (NEA). (1990). Learning laboratories initiative: Helping school districts foster local innovation. Washington, DC: Author.

Needham, N. R. (1989). Schools wait for their report cards. NEA Today, 8(5), 3.

Rothman, R. (1989). In Connecticut, moving past pencil and paper: Student assessment rates performance. Education Week, 9(1), 1,21.

Schlechty, P. C. (1976). Teaching and social behavior. Boston: Allyn & Bacon.

Schlechty, P. C. (1990). Purpose, vision, and structure: Leadership imperatives for school reform. San Francisco: Jossey-Bass.

Schlechty, P. C., Ingwerson, D. W., & Brooks, T. I. (1988). Inventing professional development schools. Educational Leadership, 46(3), 28-31.

Sizer, T. (1989). Diverse practice, shared ideas: The essential school. In H. J. Walberg & J. J. Lane (Eds.) Organizing for learning: Toward the 21st century. Reston, VA: National Association of Secondary School Principals.

Sykes, G. & Elmore R. F. (1989), Making schools manageable. In J. Hannaway and R. Crowson (Eds.) The politics of reforming school administration. New York: The Falmer Press.

Tittle, C., Kelly-Benjamin, K., & Sacks, J. (forthcoming). The construction of validity and effects of large scale assessment in the schools. In R. Stake (Ed.), Effects of changes in assessment policy. Greenwich, CT: JAI Press.

Tyson-Berstein, H. (1988). A conspiracy of good intentions: America's textbook fiasco. Washington, DC: Council on Basic Education.

Viadero, D. (1990). AFT to develop three professional-practice schools. Education Week, January 31.

Wiggins, G. (1990). 'Standards' should mean 'qualities' not 'quantities'. Education Week, January 24, 36, 2S.

Wilson, E. K. (1971). Sociology: Rules, roles, and relationships. Homewood, IL: Dorsey.