DOCUMENT RESUME

ED 374 551 EA 026 129

AUTHOR Corbett, H. Dickson; Wilson, Bruce L.
TITLE Statewide Testing and Local Improvement: An Oxymoron?
INSTITUTION Research for Better Schools, Inc., Philadelphia, Pa.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Feb 89
NOTE 34p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Standards; Elementary Secondary Education; *Minimum Competency Testing; Performance; Politics of Education; *Standardized Tests; State Curriculum Guides; *State Standards; Statewide Planning; *Teacher Attitudes
IDENTIFIERS *Maryland; *Pennsylvania

ABSTRACT

A Carnegie Foundation survey, conducted in 1988, found that teachers were critical of the education reform movement in general and standardized statewide testing in particular. This paper presents findings of a study that examined the impact of state-mandated testing programs on the work lives of teachers and students. It compares two states' testing programs--Pennsylvania's program with "low stakes" consequences attached to student performance, and Maryland's "high stakes" program. Data were collected in 3 phases: (1) interviews with administrators, teachers, and students at 12 sites (6 school districts in each of the states); (2) a survey of the central office administrator, principal, and teachers from 207 Pennsylvania districts and 23 Maryland districts; and (3) followup fieldwork. In each state, teachers perceived that the statewide testing programs offered relatively few benefits for students, particularly because they provided information that schools already possessed. Educators in Pennsylvania districts reported that they began to take the tests more seriously for political reasons but had reservations about whether the tests actually improved the lives of teachers or students. Under high stakes conditions, the following occurred: increased attention to improving test results; greater disruption of teacher' work lives; decreased reliance on teachers' professional judgment; and heightened concern about liability. Despite such programs' questionable educational value, they are politically popular because they are publicly available and easily understood. However, the press for uniform, quick success contradicts the nature of the school-improvement process. Three tables are included. Contains 18 references. (LMI)

Statewide Testing and Local Improvement: An Oxymoron?

H. Dickson Corbett

Bruce L. Wilson

February, 1989

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

The work upon which this chapter is based was funded by the Office of Educational Research and Improvement, U.S. Department of Education. The opinions expressed do not necessarily reflect the position or policy of the Department, and no official endorsement should be inferred.

STATEWIDE TESTING AND LOCAL IMPROVEMENT: AN OXYMORON?

During the early part of this decade, state departments of education, state legislatures, and governors have initiated a series of reforms designed to improve the quality of education. As an example, close to 60 percent of the states have mandated some form of standardized testing for local school systems (Marshall, 1987). Yet the effects of implementing such testing programs on the daily lives of school staff and students and how differences in state programs magnify or minimize these effects have not been well documented by empirical research despite this flurry of effort (Airasian, 1987; Rosenholtz, 1987; Stake, Bettridge, Metzer & Switzer, 1987).

A Carnegie Foundation (1988:1) survey reported teachers to be very critical of the reform movement in general and statewide testing in particular, pointing out that "The relationship between the teacher and the student is at the heart of education, and only when improvements reach the classroom will excellence be achieved." Carnegie's survey of 13,500 teachers, published in Report Card on School Reform: The Teachers Speak, found that teachers do not believe the majority of the reforms have done much positively for the classroom and are troubled by the potential for negative impacts. Concerning standardized testing, teachers noted a dramatic increase in their use over the past five years, and practitioners' comments on this development led Carnegie (1988:5-6) to conclude that "...there is something troubling - even paradoxical about these findings. We are disturbed that testing instruments are crude and often measure that which matters least;" and compounding the problem is that "In the end, what we test is what we teach." Nevertheless, the tests have been implemented with scant attention to their impact, positive or negative, on teachers and students on a daily basis.

This chapter is based on a study conducted by the authors that addressed this imbalance by documenting the impact of state mandated testing programs on the work lives of teachers and students. The chapter contrasts two states' testing programs, one with "low stakes" consequences attached to student performance and the other representing a "high stakes" situation. After comparing the programs and describing the research design, we detail teachers' perceptions of the tests' effects on their work lives and their students. Then we document differences in impacts between the low and high stakes conditions. During the research, the tests' stakes increased (dramatically in the low stakes situation) and the effects of those changes are presented next. The chapter concludes with a discussion of the value of state minimum competency testing for improvement of practice in local districts.

THE TESTING PROGRAMS IN TWO STATES

Educators from Pennsylvania and Maryland participated in the study. The states represented "low stakes" (Pennsylvania) and "high stakes" (Maryland) situations. The level of the stakes associated with a test is the extent to which students, teachers, administrators, and/or parents perceive test performance to be "used to make important decisions that immediately and directly affect them" (Madaus, 1988). Relatively minor consequences attended student performance on Pennsylvania's minimum competency tests (MCT) in language and math. The purpose of both tests was to identify students needing additional classroom instruction who may have been overlooked by other means. Maryland's "high stakes" strategy required students to pass reading, writing, math, and citizenship MCTs in order to receive a high school diploma. The tests were being phased in as graduation requirements; at the time of the survey phase of the research only the reading and math tests "counted."

The two states' MCT programs had several important differences (see Table 1). The first difference concerned the purposes detailed above. Second, Pennsylvania students were tested in the third, fifth, and eighth grades. Maryland tested students beginning in ninth grade, with a practice instrument administered in the eighth grade. Third, Pennsylvania students took tests in reading and math whereas Maryland students also were examined in writing and citizenship. Fourth, the Pennsylvania legislature appropriated funds for remediation; Maryland offered no special financial assistance. Fifth, Pennsylvania's legislated program responded to calls for educational reform in the early 1980s and, after soliciting educators' input on appropriate test objectives, invited commercial test publishers to bid on a contract to develop the state's instrument. Maryland initiated a statewide curriculum improvement program several years prior to beginning the testing program with the expressed purpose of anticipating the instructional quality necessary to perform well on the tests. Educators from around the state were used by the state department to provide input into the content and form of the tests.

Table 1 about here

The programs' stakes changed during the study. In Pennsylvania, the Chief State School Officer (CSSO) released district rankings based on the test scores prior to the 1987-88 school year and touted the test as an appropriate indicator of school effectiveness. Study interviews conducted subsequent to this event revealed considerable concern on the part of local educators that the tests were being used in ways for which they were not originally intended, even though the rankings were quickly withdrawn due to the furor surrounding them. Regardless, the importance of the tests increased for both educators

and the public. Maryland had no similar dramatic event; instead, its districts had to reconcile themselves to the inevitable day when all four tests would affect whether students graduated, with the writing and citizenship tests generating much controversy and calls for revision. In fact, administrators and teachers reported that students had difficulty passing these two tests and that this augmented the pressure on them.

Madaus (1988) and Airasian (1987) argue that such differences in the purposes of statewide testing programs should have different impacts on their respective school districts. Because high stakes tests are used for important decisions such as promotion or graduation, they have the ability to influence system behavior--even to direct it (Madaus, 1988). In low stakes situations, no important sanctions follow test performance and thus the tests likely would have little effect on the system. Airasian (1987) claims that standardized testing once served general purposes, namely to identify areas where instruction needed improvement and to gauge how well the educational system as a whole was functioning. More recently, the success of these traditional uses of the tests has led to acceptance of a new purpose.

> This second use is most aptly termed state-mandated certification testing. In this approach, testing is not used to guide classroom instruction or to monitor educational policy. Rather, state-mandated certification testing has made testing and test results a crucial aspect of educational policy itself. (Airasian, 1988:403)

In other words, states have begun to use tests as the policy to try to spur improvements. These tests, of which MCT is one form, often have common characteristics: they are mandated for most students in selected grades; they eliminate local discretion by using one instrument to be administered and scored similarly across all systems; and they usually measure performance on a pass/fail basis. The consequences of such a testing policy are that test

information becomes of interest to a wide population and not just a few professionals and concerned parents, local control over the curriculum may be eroded, and a tension is created between quality of education and equality of educational opportunity.

The two states examined in this chapter were selected with this potential for differences in impact in mind. Pennsylvania's approach was much more in line with Airasian's (1987) assessment of the traditional use of standardized testing and contained low stakes for the system as whole, although remediation money was given to districts on the basis of how many students fell below the cutoff point--a potential negative incentive for improving scores. Maryland, on the other hand, designed the test as a specific policy tool and tied student performance to high school graduation. Thus, the administrators, teachers, and students in this state faced a high stakes situation.

The available literature offers little guidance as to what precisely the differential impacts of such programs might be. Stake et al. (1987) provides an initial review of research on the effects of state assessment initiatives, examining the topic across six categories of effects: achievement standards; public attitude toward schools; the morale/motivation of those tested; the utility of test information for school administration; the reactions of teachers to standardized test results; and the curriculum. The review notes that few studies have been conducted to compare the local system consequences of statewide standardized (and/or minimum competency) testing programs.

Airasian (1987: 408) in a review of testing research suggests: "The crucial issues of testing are not technical. Issues of testing today are social, economic, and value-laden, involving the distribution and redistri-

bution of resources and prerogatives". Research on minimum competency testing, defined in policy terms as "a device for conditioning student promotion or graduation on test achievement" (Darling-Hammond & Wise, 1985:318), has not yet caught up with this argument.

There are several reasons why higher stakes situations can be expected to have greater local impacts. First, mandatory tests are likely to force adjustments in a system by creating expectations for what the outcomes of schooling should be. According to Mintzberg (1983), stipulating outcomes is one means used widely in organizations to affect operations. Some standard--no matter how narrowly defined--has to be met, regardless of what else staff members may want to accomplish. In situations where the standard is easily attained, its importance as a criterion of success may remain no more preeminent than any of a myriad of indicators. However, in situations where the standard is less readily reached, its importance looms larger and perhaps more directly defines what happens in the schools.

Second, one of schools' primary tasks is to move students smoothly through a series of grades to graduation (Schlechty, 1976). Staff size, the number of classrooms needed, and the availability of sufficient materials are all predicated in most communities on the assumption that essentially all first graders will become second graders and that most seniors will graduate on time. A few exceptions cause no problems, but testing programs change the assumptions by inserting a checkpoint for determining the progress of all students, based on something other than student age, credits obtained, or time spent in school. Obviously, some checkpoints are more formidable than others, as in the case where successful completion of the test determines whether or not students graduate. But even relatively innocuous checkpoints may force

some remediation and thereby affect subsequent progress.

Third, establishing a standard all students must meet as a visible indicator of effectiveness runs counter to the ethos of many educators (Rosenholtz, 1987). In spite of enormous standardization, a tone of individualism permeates American education (Lortie, 1975). Teachers are allowed considerable autonomy in deciding what and how to teach, and they expect to handle their classrooms themselves. Testing programs challenge this ethos. Test items highlight critical content to cover; test administration dates determine the deadline for teaching the content; item formats affect how the information will be accessed; and the standards add a quality of sameness to what students should achieve. The tests, therefore, have major effects on school culture. Wilson (1971) defines culture as "definitions of what is and what ought to be . . ." Deal (1985) describes it as "the way we do things around here." Testing programs are likely to require serious examination of definitions of what being a student or teacher is and should be. The literature on educational change is replete--although this is not always recognized--with descriptions of the clash between values implicit in an innovation and the values implicit in the way those expected to innovate were accustomed to behaving (Sarason, 1971; Gordon, 1984; Rossman, Corbett, & Firestone, 1988).

Of course, greater impact is not tantamount to improvement. After describing the study design, the remainder of the chapter will be devoted to detailing the type of impacts local systems felt with respect to teachers' and students' work lives.

STUDY DESIGN

The above discussion simplifies a complex situation. Introducing and operating a mandatory statewide MCT program involves a wide range of potential challenges to a district. While some of these challenges can be anticipated by theoretical understanding or past research, using an inductive approach in which the present research can take advantage of unexpected developments can be equally valuable (Miles & Huberman, 1984). For this reason, the study was designed to include both in-depth, open-ended qualitative fieldwork in a small number of sites and large-scale structured questionnaires.

The study had three phases. First, researchers conducted a preliminary round of qualitative fieldwork wherein they visited six school districts in each of the two states for several days to interview a wide variety of staff members. Second, the results from the interviews were used to design a questionnaire to be administered throughout districts in the states studied. Third, the survey results were used to structure a final round of feedback and interviews in the original sites. These latter interviews were conducted with mostly administrators during half-day visits in 11 of the 12 districts.

Phase One: Fieldwork in 12 sites

Six sites in each of the two states were visited. Site selection was made on the basis of district size and type of community served, primarily because these characteristics were assumed to determine the kind of staff resource demands providing test-related followup instruction would take. Equally important was the willingness of the district to participate because the purpose of this phase was to explore issues in depth, not to generalize to a larger population. Selection was carried out with the input and assistance of key state department staff members in each state.

Six experienced field researchers conducted the site visits. One researcher spent two or three days in each site depending on district size. The first day was spent in the central office, interviewing the superintendent (if available), the person(s) responsible for handling the testing program, and other district staff members who dealt with the test. Also, pertinent documents were examined where available. On days two and three, school interviews were conducted with administrators, guidance counselors, teachers, and students. When all appropriate schools in a district could not be visited, selection was made in collaboration with district personnel. Sampling a variety of schools in the district was the foremost criterion. Over 250 local educators and students participated in the interviews.

Interview Questions. Field researchers operated from interview guides with broad categories of questions. For further documentation of interview protocols and data summaries the reader is referred to Corbett and Wilson (1988). Specific phrasing of questions and the particular probes used were determined by the researcher on site. In training sessions conducted prior to the site visit, researchers had an opportunity to generate and discuss potential questions and follow-up probes, but fieldwork of this type demands that the researcher have considerable flexibility in determining who to talk to, what to ask, and when to ask it. The goal was to obtain data on each question from multiple sources but not necessarily from every source.

Data Management. A multiple-case, multiple-researcher, open-ended interview study places a heavy burden on the data management system. A systematic way of determining data gaps, locating overlooked sources, making data accessible to other researchers, and being able to retrieve parts of the data was imperative. To accomplish this, resources were allocated more to

developing data summaries than to making handwritten field notes presentable or typing transcripts from tape recordings. When researchers returned from a site visit, they completed a series of data summary charts: (1) a summary of information sources and the question categories for which each source supplied information; (2) a description of source-identified effects coupled with the researcher's designation of which and how many staff members listed each effect; (3) a summary of data on the district's instructional, organizational, and cultural contexts as well as its relationship with the surrounding community and the state; and (4) a listing of residual incidents and data worthy of note that did not fit cleanly in the structured charts.

These data summary charts were used by the authors to conduct the cross-site analysis. They were the stimulus for determining whether additional information needed to be gathered from particular sites.

Data Analysis. The analysis activities consisted of reviewing the data summary charts to identify implementation themes that cut across the 12 sites. The specific goal of the analysis was to develop items for the questionnaire to be used in the second phase of the study.

The authors returned to the original field notes to review the terminology local educators used in discussing the tests. Using the list of themes, the data summary chart information, and this review of responses, individual questionnaire items were constructed. A questionnaire with 83 items was produced from this synthesis. The items fell into five categories: local internal and external operating contexts; the administration of the tests in the local setting; the strategies used to maximize student performance; the purposes the tests were used for in the local setting; and the impact of the tests on instruction, organization, and culture.

Phase Two: Survey Design

The second phase of the study involved a quantitative assessment of the local ramifications of mandatory statewide testing programs. Four major activities--instrumentation, sampling, data collection, and analysis--were conducted during this phase.

A first draft of the questionnaire was designed so that it could be self-administered in 20 to 30 minutes. A pilot test of the draft instrument was administered in several districts to ensure that the questionnaire was clear, communicated the intent of the project, and could be completed within time constraints. Changes to the questionnaire were made on the basis of the criticism that was offered.

All districts in both states were invited to participate in the study (Pennsylvania = 501; Maryland = 24). Three different role groups familiar with the testing program were targeted from each district: central office administrators, principals, and teachers. A separate questionnaire was completed by each role group member. In Maryland, where there were fewer but larger school districts, three respondents from each role group within the district were asked to complete the survey. Only one person from each role group within the district completed the survey in Pennsylvania. The participating staff members in each system were selected by the superintendent or a designee.

In Pennsylvania, 277 of the 501 districts responded with one respondent from each of three role groups (central office, principal, and teacher). In Maryland, 23 of the 24 districts returned useable questionnaires with three respondents from each of three role groups. An analysis of the participating

and non-participating districts in Pennsylvania showed no significant differences between the two groups in terms of basic demographic characteristics (e.g. size, wealth, location).

The analysis had two foci. The first was to identify educators' responses concerning the adjustments they had made. Frequency distributions for questionnaire items were used to display these responses. The second focus was to examine cross-state differences for instructional adjustments. Analyses of variance were conducted to compare responses in the two states.

Phase Three: Follow-up Fieldwork

In the fall of 1987, field researchers returned to 11 of the original 12 sites visited in Phase One, with one Maryland district declining to participate. The purposes of these visits were to trace subsequent developments in the operation of the state testing program and to obtain assistance in interpreting the results of the survey. Over 80 local educators participated in this activity. The interviews concentrated on the findings contained in the section on within-state district variations. The findings were presented to participants and they then reacted to specific numbers, interpretations, and implications. These reactions then were incorporated into the quantitative results section of this chapter.

FINDINGS REGARDING EDUCATORS' REACTIONS TO STATEWIDE TESTS

This section gives a flavor of how educators felt about their respective state's program and hints at important differences between the two states. The specific focus for this chapter is on items addressing teacher work lives and the lives of students. In each case, sample items representing the general theme of teacher work life and student life were included in the survey.

The "Student Life" items were not intended to comprise an all encompassing category. The items included in this group offered a glimpse of how the character of student life fared under the testing program in terms of the extent of change in each of the following areas:

- Students are more serious about their classes.
- Teachers have more empathy for students who are achieving poorly.
- Staff members know more about students who have serious learning problems.

Respondents were asked to indicate the extent of impact on a five point scale (0-4) from "no change" to "total change" as a result of the testing program. A higher score meant a more positive impact on student lives.

Similarly, the "Teacher Worklife" category sampled six items focusing on the extent of change in important conditions that define the working conditions for teachers, such as:

- There is a decreased emphasis on using educators' professional judgment in instructional matters.
- Time demands on staff have increased.
- Staff members have been reassigned.
- Staff members are under pressure to improve student performance.
- Paperwork has increased for staff.
- Staff members are more worried about the potential of a lawsuit.

This measure was not intended to be inclusive of all aspects of work generally discussed in the working conditions literature, but at least the items provided an indication of whether teachers' work lives were affected by the new testing programs. Like the previous items, the respondent choices were on a five point scale (0-4) from no change to total change. In this case, the higher the score, the more stressful the work environment for teachers.

Frequency distributions for the respondents in each state are presented in Table 2. Teacher responses were used rather than responses from the other two role groups surveyed (building principals and central office administrators) because it was felt that teachers were in a better position to be informants about their own work lives and student lives than other role groups. In each district the teachers(s) were nominated by the superintendent because of their knowledge of and experience with the state testing program. The numbers in the table represent the percent of teachers responding to each category.

Table 2 about here

The findings from the three items focusing on the quality of student life indicate that teachers, on average, were reporting only minimal impact as a result of the test. Approximately half of the teachers from both states reported "no change" or only "minor change" on students as a result of the test. As interviewees commented:

> The students are not impacted. The test identifies the same kids with the same problems [as other diagnostic instruments]. No one had to tell us who was having problems. They had already been identified.

> The student impact is low because there is nothing obstrusive to affect students.

The mean scores for the three items demonstrate further the minimal impact but also reveal some differences between the two states. On average, the impact is "minor" in Pennsylvania and between "minor" and "moderate" in Maryland. Although differences are present, there are no substantial added benefits for students, particularly in whether teachers know more about students than before. This is interesting in that improving this knowledge is a common

justification for beginning the testing program in the first place. Apparently teachers feel confident that they do not need additional tests to show them who needs help.

Another explanation for the low positive impact on students may be the counterbalancing negative impact that teachers mentioned during interviews:

> Testing adds a negative image for kids who fail. It's another way of telling someone "I'm dumb." It makes it difficult for kids to get up in the morning.
>
> Those who fail are second class citizens...we take them out of regular instruction for remediation.
>
> We don't get the mileage out of better kids that we used to . We are teaching to the middle.

The findings from the six items sampling the quality of teacher work life reveal a greater relative impact in Maryland, the high stakes site. For three of the six items - time demands, pressure for performance, and paperwork - over half the respondents in Maryland indicated the change was "major" or "total". Comments in the interviews reinforced the negative impact on teachers lives:

> Teacher self-esteem goes down another notch each time something like this happens.
>
> The paperwork is horrible and getting worse.
>
> Professionals aren't trusted -- the tests carry the aura of respectability.
>
> It takes too much time...too much of that time has to be taken from other stuff I used to do.
>
> Teachers feel jerked around. The test tells them what to teach.

On average, Pennsylvania teachers reported only "minor" change ($x = 1.0$) while Maryland teachers reported the impact to be between moderate ($x = 2.0$) and

major ($x = 3.0$). Clearly, the differences in high and low stakes conditions accentuated the impact of testing on teachers' work lives.

The findings reported in Table 2 offer a snapshot of teachers' reactions to the initiation of statewide mandatory minimum competency tests. The survey findings suggest that, regardless of stakes, teachers believe relatively minor benefits flow to students. The evidence is stronger when considering teacher work lives. Teachers perceive the tests as placing more negative demands on their already overcrowded schedule: "The test is just one more add-on activity."

STATE COMPARISONS

Clearly, for the reasons discussed earlier, Maryland's MCT program should have had a greater impact on its local systems than Pennsylvania's program, primarily because Maryland's policy insinuated itself into an important organizational event--graduation--and because preceding statewide improvement and actual test development activities engendered a cumulative anticipation of the day the tests would be put into place. On the other hand, Pennsylvania's program arose from dialogue limited mostly to state level legislators and officials. Limited local knowledge about the program plus its lack of implications for school operations seemed to insure that the test would have little impact beyond its stated purpose of identifying students in need of additional instruction.

The results in Table 3 assess the differences between teacher respondents in the two states. A mean score for each respondent was computed by combining the three student life items into one scale and the six teacher work life

items into another. An analysis of variance was conducted on the two scales. Prior to combining these items to create a scale, statistical tests were conducted to ensure the appropriateness of such a step. First, correlation matrices were examined to insure that there was at least a moderate correlation among the combined items and that there were not any excessively high correlations. Second, an analysis of reliability (internal consistency) was conducted to test that the items cohered together. The results of those calculations produced a reliability coefficient of .70 for student life and .83 for teacher work life, suggesting high internal consistency.

Table 3 about here

The findings were striking and consistent. For both measures, statistically significant differences between the states were found. Staff in Maryland school districts reported more impact on students and their own work lives than their Pennsylvania colleagues.

Essentially, the two states had different intentions in mind when the testing programs were initiated and the study data indicate that both were being met. The data reflect the differences in the modest versus the more ambitious approaches.

Recent Developments in the Two States: Raising the Stakes

The above quantitative comparisons present a snapshot of the differences in teachers' reactions to the testing programs. The picture was taken in the late Fall of 1986 and the early Winter of 1987. Events in both states subsequent to the administration of the survey seemed to increase the level of the stakes associated with the tests and had an effect on staff sufficient to

alter their perceptions of the effects of the testing programs. In both states, an increase in the impacts on students and teachers were noted. A detailed account of these changes is available in Corbett and Wilson (1988).

The key event in Pennsylvania was the publication of the results from the Spring of 1987 test administration. Rather than the customary low-key sending of the scores to districts for each to handle as it saw fit, the release of the data was orchestrated by the chief state school officer (CSSO). In a public media briefing, the CSSO provided documents that ranked districts in the state from top to bottom in terms of the percentage of students who passed the cut-off point on the MCT. In addition, schools that had achieved 100 percent passing rates despite having "high risk" student populations were singled out as being "poised on the brink of excellence." And to cap off the presentation, the CSSO touted the tests as the best measure available to assess the effectiveness of Pennsylvania's schools. An immediate protest over this use of the scores arose from educators across the state and resulted in the withdrawal of the documents containing the rankings.

The withdrawal of the rankings did not strike the event from either educators' or their communities' emotional record. Administrators in three of the six Pennsylvania districts visited in Phase Three argued that the "game" had now changed in their systems. Reflecting on the impact on students and teachers, they commented:

> The purpose of the test changed in September. It is no longer for remediation but to rank order schools.
>
> The results should be between the state and the school district if the test is to help. When they release scores and say 58 kids need help, we can say we've already identified 40 of them. But the negativism starts; it starts [phone] calls and there is no question I now have pressure on me.
>
> The test was not all that important....But we might as well face up

> to it; with the publication of school by school results....one of the goals will be to raise the percentage above the cut score.

What really seemed to be changing for the three districts in Pennsylvania were the stakes; they got higher, primarily through the increased visibility of score comparisons and the subsequent increased, albeit reluctant, acceptance of the scores as a benchmark--that is, as a widely recognized point of reference when discussing the performance of schools in the district and in surrounding districts. Staff in the three districts reported that they did not believe the tests to be particularly important educationally and did not embrace the tests as valid indicators of achievement. They nevertheless acknowledged that they already were, or would soon be, treating the scores more seriously than in previous years.

This is best illustrated by a district whose surrounding districts performed similarly on the MCT, even though the district felt that its carefully and systematically developed curriculum far surpassed the offerings of their neighbors. The response from the superintendent:

> We don't believe in the tests that strongly but we will be forced to see all material is covered before the tests. We definitely are going to do it. We won't be caught in the newspapers again.

The brunt of not "getting caught" again was to be borne by the reading program--a recently revised, developmental curriculum. The timing of the test administration required shifting the sequence of topics to be covered. An outraged reading coordinator responded,

> You have to alter a curriculum that is already working well and so we can't follow the developmental process. Kids are already growing in a structured program; but it [pressure to change] comes from the board, community, and adverse publicity.

The superintendent empathized with the coordinator,

> I don't have much faith in the tests. I don't want to change the curriculum, and it's not a major revision, but we've got to do better. Still, it's not the right thing to do to anyone. I don't

> want to over-react but I'm also going to have to spend time on things I shouldn't have to do as well: public relations, testing meetings--just to make the board feel comfortable. It'll never happen again when we see a worse district doing better than us.

The interviews suggest that these districts were planning expedient strategies to improve the test scores and just as clearly that there was resentment to do so and a concern that what they were doing was compromising some standard of good professional practice. The message they were giving was that their test scores were becoming benchmarks for political reasons, namely to appease school boards and community members who had had the opportunity to see their school systems compared to neighboring districts and did not like what they saw.

No single event dramatically heightened the impact of the tests in Maryland. Instead, the stimulus was the approach of the time when students had to pass all four of the tests in order to receive a diploma. The four tests were not regarded equally. Phase Three interviews revealed that educators discriminated between the reading and math tests on one hand and the writing and citizenship ones on the other. The reading and math tests, in Maryland educators' minds, were adequate measures of basic competence in the respective content areas and covered objectives already well-entrenched in the curriculum. The curriculum development aspect of the state initiative began in the late seventies, and these two tests were the first to be developed, trial-tested, and implemented. Actual local curriculum and instruction changes had been in place for seven to nine years in some settings. By 1987, these alterations had become institutionalized to the point that interview subjects in four of the five Phase Three districts argued that the impact scores may have been too low because staff had forgotten that what was now

routine was once novel. The result was that the two tests were no longer intrusive.

Such was not the case for the writing and citizenship tests. Both generated considerable controversy. The writing test did so primarily because staff viewed it as demanding a performance level well beyond that necessary to be minimally competent in writing. The citizenship test was controversial because it required students to memorize information about local, state, and federal governments--information that even the teachers said they did not possess without special study. Fueling educators' concerns were the facts that students had much more difficulty succeeding in trial administrations of these two tests and that the time when the first cohort of students would have to pass all four tests to receive a diploma was inexorably approaching. For special education teachers and teachers with responsibilities in the grades tested and for affected content areas, the pressure to achieve passing scores was building and the impact on their work lives was great. According to two administrators:

> We've changed the whole social studies curriculum. We had to expand the 7th and 8th grade American Studies to include more history (to make up for content not being taught later) and now teach government in the last term of 7th and 8th grades which we did not teach at all as a separate entity in the past. And we have structured in key points in the language arts scope and sequence.
>
> It depends on who the teacher is and what the teacher teaches. You can't have a bigger impact than on sequence or inserting a new course. We now offer courses not included before and content that changed from 10th to the 9th grades. With government, the impact is overwhelming.

As illustrated in the above quote, there was a "differentiated" impact of implementing the tests. Some parts of the system were affected little while others felt considerable ramifications. Such a situation caused statistical

measures of central tendency such as the mean scores presented above to disguise this important impact of the tests.

The "discomfort" of subgroups of staff involved with the two controversial tests focused their attention more and more on the percentage of students passing the tests and on adopting expedient methods of improving scores. This "concentrated" approach, was apparent in all five systems where Phase Three interviews were conducted.

> We are concentrating more on basics. We are now spending from September to November on basic skills rather than on our developmental program. [reading teacher]
>
> I'm not opposed to the idea of testing. But I'm not so sure we haven't gone overboard, the tail is wagging the dog. The original idea was that there were to be certain standards the student would have to meet, but if the student doesn't pass, people will ask what's wrong within the school and teachers. [teacher]
>
> When the scores are low, it takes me into the school for the names of the kids who failed. There is no stroking in schools where scores have dropped. Everyone is sitting around with bated breath waiting for the test scores. [central office administrator]
>
> We realize a kid is taken out of science every other day for citizenship and will fail science to maybe pass the citizenship test. [building administrator]

These very targeted means for getting students to pass were acknowledged as a necessary evil:

> We've had to do things we didn't want to do. [central office administrator]
>
> We have materials provided by the county as 'quick help.' We were told 'here's how to get kids to pass the test fast.' They were good ideas but specifically on the test. For example, if the area in a rectangle is shaded, you multiply; if not, you add. [teacher]

And in response to the above stream of comments, a teacher summarized,

> Talk about games and game-playing!

It is important to note that the stakes were raised in the two states for two different reasons: (1) public pressure to improve test scores that

resulted from readily available comparisons of performance in Pennsylvania, and (2) the proximity of both the yearly test administration day and the day when the two troublesome tests would actually serve as an obstacle to graduation in Maryland. Interestingly, the stakes increased in what were originally both low and high stakes situations. As they did so, educators' concern shifted almost completely to influencing test performance. Put differently, the manifestations of the seriousness with which the test was taken shifted. The change can best be described as one from a long-term focus to a short-term one, from using the tests as one indicator among many to treating the next set of test results as the most important outcome of schooling.

CONCLUSION

Under either the low or high stakes condition, teachers perceived that the statewide testing programs offered relatively few benefits for students, particularly in terms of providing additional information that schools did not already possess to determine which students could be better served. There seemed to be little justification in educator's minds for adding another test to the set of existing instruments being administered at the local level simply to identify several more students in need of special instruction.

However, in high stakes situations, great attention was paid to this admittedly uninforming information. An important question is: Was this increased attention to test scores for the better? The qualitative data from Phase Three of the study suggested that as the perception of the importance of the test increased, there was a point at which district responses took on the flavor of a single-minded devotion to specific, almost "game-like" ways to

increase the test scores. Pennsylvania districts, in particular, that began to take tests more seriously reported that they did so for political reasons and not because they believed that they were actually improving the lives of students or teachers. Prior to this point, the strategies emphasized more systematic changes in the curriculum. Beyond this point, staff began to respond to questions about effects with the phrase: "Some good things have happened as a result of the tests, but..." Staff members' reservations about the practices they were engaging in to improve the scores followed the "but." When the stakes but not the quality of the information contained in the tests changed, so did local attention to improving scores. But a turning point was reached, and the modest positive effects associated with having additional diagnostic information available was overwhelmed by perversion of local practice, with the primary goal becoming to improve test scores. Many of the negative behaviors associated with "teaching to the test" thus emerged. The exact turning point likely varied from district to district; but it was clear that the test scores were beginning to govern activity more directly, as Minzberg (1983) predicted could be the case when an organizational outcome increases in importance.

Concomitant with increased attention to improving test results was greater disruption to teachers' work lives. Although teachers acknowledged that a narrowed curriculum could also be an improved one (Wilson & Corbett, in press), few indicated in interviews that their actual teaching had improved. To the contrary, they reported that they occasionally strayed from sound instructional practices in order to get students to pass. They also reported that, under high stakes conditions, there was a decreased reliance on their professional judgment in instructional matters, increased time demands, more

staff reassignments, greater pressure, more paperwork, and heightened concern about liability.

If a statewide testing program engenders little additional benefit for students and greater disruption for teachers without improving practice, then it would seem the program has little educational value. So why the popularity? Statewide tests are primarily a political device; they are easily legislated and--when results are reported in the form of passing grades--easily interpreted. These results can be effective rallying points to mobilize pressure on almost any school or system to "improve," depending upon what level of success is deemed appropriate by a particular community. Thus, the presence of publicly available and understood results affords the opportunity for greater state and local community involvement in determining what goes on in the schools.

It is interesting that both state policymakers and community members define improvement as greater standardization across schools and as achievable within a yearly testing cycle. The press for more uniformity and quick success, however, contradicts everything that is known about the process of improving schools. School improvement succeeds when the idiosyncracies of school demographics, culture, and organization are taken into account in a process that incorporates generous dollups of technical assistance and staff interaction within a three to five year time span. It takes considerable time to plan what to improve, to try out means of attaining that goal, to assess which means are effective, and to take steps to insure that the effective means become part of the operational routine. The testing cycle, on the other hand, forces the compression of this process into a single year and increases reliance on the lifeline of a common set of testing objectives (regardless of

the student population) to avoid drowning in a sea of public criticism. The only available escape is to focus directly on those objectives, a strategy that definitely can raise test results in the short term but accomplishes little systemic improvement in the long term.

REFERENCES

Airasian, P.W. (1987). State mandated testing and educational reform: Context and consequences. American Journal of Education, 95(3), 393-412.

Carnegie Foundation for the Advancement of Teaching (1988). Report card on school reform: The teachers speak. Princeton: Author.

Corbett, H. D., & Wilson, B. L. (1988). Raising the stakes on statewide mandatory testing programs. In R. Crowson & J. Hannaway (Eds.), The politics of reforming school administration. Falmer Press: Philadelphia.

Darling-Hammond, L., & Wise, A. (1985). Beyond standardization: State standards and school improvement. Elementary School Journal, 85(3), 315-336.

Deal, T. E. (1985). The symbolism of effective schools. The Elementary School Journal, 85(3), 60-620.

Gordon, D. (1984). The myths of school self-renewal. New York: Teachers College Press.

Lortie, D. (1975). Schoolteacher: A sociological study. Chicago: University of Chicago Press.

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

Marshall, J.C. (1987). State initiatives in minimum competency testing for students. Policy Issue Series No. 3. Bloomington, IN: Consortium on Educational Policy Studies.

Miles, M. B., & Huberman, A. M. (1984). Qualitative data analysis: A sourcebook for new methods. Beverly Hills, CA: Sage

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

Rosenholtz, S.J. (1987). Education reform strategies: Will they increase teacher commitment? American Journal of Education, 95(4), 534-562.

Rossman, G. B., Corbett, H. D., & Firestone, W. A. (1988). Change and Effectivenss: A cultural perspective. Albany, NY: SUNY Press.

Sarason, S.B. (1971). The culture of the school and the problem of change. Boston: Allyn and Bacon.

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

Stake, R. E., Bettridge, J., Metzer, D., & Switzer, D. (1987). Review of literature on effects of achievement testing. Champaign, IL: Center for Instructional Research and Curriculum Evaluation.

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

Wilson, B. L. & Corbett, H. D. (in press). Two state minimum competency testing programs and their effects on curriculum and instruction. In R.E. Stake (ed.), Advances in Program Evaluation: Effects of changes in assessment policy, Volume 1. Greenwich, CT: JAI Press.

TABLE 1

Summary of Two Mandatory, Minimum Competency,
State Testing Programs in Pennsylvania and Maryland

Areas of Difference	Pennsylvania	Maryland
TEST CONTENT	Reading, Math	Reading, Math, Writing, Citizenship
GRADES TESTED	3, 5, 8	8 (Practice) 9, 10-12 Retests
PARTICIPATION	Mandatory	Mandatory
STATE FOCUS	Use of test results to identify students in need of additional instruction	Identification of failing students to aid districts in curriculum planning
LOCAL CONSEQUENCES	Additional state funds for low scoring students	Students must pass test to graduate; districts required to provide appropriate assistance to failing students; no additional state funds
STIMULUS	Legislative response to reform based on critiques of early 1980s	State department curriculum improvement initiative begun in late 1970s

TABLE 2
Percent of Respondents by State for Student Life and Teacher Work Life Items

Student Life	No Change		Minor Change		Moderate Change		Major Change		Total Change		Mean Score	
	PA	MD	PA	MD	PA	MD	PA	MD	PA	MD	PA	MD
1. students more serious	37	26	32	34	25	28	5	12	0	0	1.00	1.26
2. increased empathy for poor achievers	38	12	27	35	30	38	5	13	0	0	1.02	1.58
3. more known about learning problems	18	15	27	35	42	33	13	16	1	2	1.54	1.56

Teacher Worklife	No Change		Minor Change		Moderate Change		Major Change		Total Change		Mean Score	
	PA	MD	PA	MD	PA	MD	PA	MD	PA	MD	PA	MD
1. decrease in professional judgment over instructional matters	63	26	22	16	12	34	4	20	0	4	0.56	1.60
2. increased time demands	39	3	25	0	25	27	9	52	2	18	1.10	2.81
3. staff reassignment	58	15	23	31	14	37	3	12	3	6	0.70	1.61
4. increased pressure for student performance	36	3	26	8	24	25	12	46	2	18	1.17	2.67
5. increased paperwork	41	2	29	2	21	31	6	43	3	23	1.01	2.84
6. more worry about lawsuits	74	19	13	37	9	22	4	15	0	7	0.43	1.56

NOTE: The number of respondents varies from the number reported in the discussion of the research design because a few districts did not provide teacher respondents and because some respondents did not answer all the items. The average number of valid responses in Pennsylvania was 250 and 57 in Maryland.

TABLE 3
Analysis of Variance Comparison of Student Life
and Teacher Work Life Scores by State

Cluster	Mean PA	Mean MD	F
Student Life	1.29	1.48	6.3**
Teacher Work Life	0.81	2.17	152.2***

NOTE: The number of respondents varies from the number reported in the discussion of the research design because a few districts did not provide teacher respondents and because some respondents did not answer all the items. The average number of valid responses in Pennsylvania was 250 and 57 in Maryland.

** $P \leq .01$
*** $P \leq .001$