DOCUMENT RESUME

ED 085 746 CS 200 926

AUTHOR Diederich, Paul B.
TITLE What Statewide Testing Can Do.
PUB DATE Nov 71
NOTE 12p.; Paper presented at the Annual Meeting of the National Council of Teachers of English (61st, Las Vegas, November 25-27, 1971)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Curriculum Evaluation; Education; *Educational Assessment; Evaluation; Learning; *State Surveys; *Testing; *Test Interpretation; Test Results

ABSTRACT

Statewide testing can serve four important functions: can illustrate superior results of a group of schools where no one would expect it and raise questions about how they accomplished it; statewide testing deals with the generally lower scores of disadvantaged minorities, it can put the differences in perspective by showing comparable differences between boys and girls; testing statewide can deal with school effects other than knowledge and basic skills, as shown not only by an interest measure but also by data on attitudes toward school; and it can show that a particular program is producing substantial and socially important results. It is not necessary to give the same test to everybody in the whole state if the objective is to discover the strong and weak points in the state's educational system. (WR)

WHAT STATEWIDE TESTING CAN DO

Paul B. Diederich
Educational Testing Service
Princeton, New Jersey

Instead of discussing theoretical possibilities of statewide testing or its present character and extent, I want to show some results of more than local interest, embodied in four exhibits that I hope you have received. Some were based on an area larger than a state, but these kinds of information can be secured in any state testing program.

The first graph, entitled "Selected Indices for Towns in Two Regions of One State," does represent a single state. The subtitle says "Grade 4." We tested everybody in grades 4 and 7, but the results were so nearly the same that I saw no need for a second graph. This state was divided into five regions with different characteristics, and I chose towns of 2,500 to 10,000 population to represent two of them because both regions had a lot of small towns but not comparable numbers of anything else. Neither included a metropolis or its suburbs. White and shaded bars show where these towns stood in relation to all 600 school districts in the state, including many large communities.

Note first that white towns have more money. They stand at the 80th percentile in the socio-economic status of their pupils; shaded towns at the 50th. But the shaded towns are 30 percentiles more favorable toward school in the attitudes expressed by their pupils in unsigned questionnaires. Right away we have data on something other than knowledge and basic skills.

In spite of lower instructional costs per pupil, shaded towns have slightly more teachers with Master's degrees.

The big surprise is that shaded towns stand at the 90th percentile of all school districts in the state in vocabulary, which was used as a rough index of brightness, and in "composite achievement"--reading, writing, and arithmetic. White towns in the wealthier region stand at or near the 40th.

This difference was not due to race. Neither region has a serious minority problem, and the white towns are higher in the socio-economic status of their pupils, which would hardly be true if they had a large Black or Chicano population. How did the shaded towns manage it? At present no one knows, but the obvious next step in this program is to send people into these towns to find out how the less affluent made a little money go such a long way.

Strangely enough, this was not done. The legislature had already decided to award compensatory state aid to the districts that did not do well on these tests in order to bring them more nearly up to par. Hence the white towns, which already have more money and higher school costs, will get even more, while the shaded towns may lose some of the state aid they have received in the past on less sophisticated bases, such as property valuation and percent of families on relief.

Two things can be said in extenuation of this action of the legislature. First, it took sharp eyes to discover that small towns in the poorest region were doing remarkably well. I had to dig this information out of widely separated graphs with so many lines on them that it was hard to see what was happening. It was not what anyone would expect. How many of you would expect small towns in the poorest region of your state to stand near the top in school achievement? This finding runs counter to most evidence of school achievement that has been gathered in the past, and we still do not know what to make of it.

The second defense is that the policy of using state or federal funds to overcome the weak spots in our educational program rather than as a reward for excellence is in general sound. I would not argue that the towns represented by the shaded bars should get a lot of state aid as a reward for doing well. For all I know, that might spoil them. But I would argue that they ought not to lose anything and that some of the money allocated to state testing programs should be reserved for investigation of unexpected successes.

As a testing man, I hate to have people think of a testing program as a way to find out what they are doing wrong. It can reveal problems, but it can also reveal what some people are doing right. As we find out what accounts for their success and promote its adoption by judicious publicity, it will go a long way toward overcoming our mistakes.

My second figure bears the provocative title, "Are Boys Genetically Inferior to Girls? or Negroes to Whites?" Look first at the graph on the right, which gives the usual picture of Negroes lagging behind whites in reading in both academic and nonacademic curricula. This graph is a bit more credible than most because it came from a large ETS longitudinal study in which the same students were tested as they advanced through grades 5, 7, 9, and 11, and those who dropped out before grade 11 were eliminated all the way back to grade 5. Hence the picture of growth in reading is not falsified by the disappearance of the poorest readers between grades 9 and 11. It is hard to estimate how many tons of ink have been spent on pictures like this, trying to prove that Blacks are or are not inferior, and that schools are or are not to blame.

The saving grace of a large study like a state testing program is that there is always more than one graph, and I found another, drawn to the same scale, in a different chapter written by a different investigator, showing the relative positions of boys and girls in reading, also in academic and nonacademic curricula. At once it is obvious that boys lag behind girls in reading in both curricula to almost the same extent that Blacks lag behind whites, but nobody writes a book to prove that boys are or are not genetically inferior to their sisters, or that schools are or are not to blame. We accept these differences as the normal result of differences in interests, expectations, and life styles. In my own family my daughter always had her nose in a book while my son never read anything but comic books as a last resort on rainy days.

Of course their scores differed on reading tests, but nobody worried about it--even in a house in which there were about two thousand books. I wish we could accept with the same equanimity the lower reading scores of Black boys and girls, who often live in apartments in which there are no books at all.

It is true that the distances between Blacks and whites in the graph are slightly larger than the distances between boys and girls. This apparent difference may not be significant, but even if it is real, it is a natural result of the fact that these boys and girls, tested in the same schools, were mainly brothers and sisters, growing up in the same homes and neighborhoods, while the whites and Blacks came from homes and neighborhoods that were terribly and often tragically different. It seems wonderful to me that the likenesses between these two graphs are much more obvious than their differences.

The point of this little demonstration is that, if your state testing program devotes any attention to the lower scores of disadvantaged minorities, I hope you will insist that they keep these differences in perspective by showing the comparable differences between boys and girls. And remember that the boys are not inferior in everything. When it comes to math and science, the position of the two sexes is generally reversed.

My third exhibit (on the back of the page) is a table comparing the interests of 11th grade boys and girls in twelve school subjects. I put in this table to illustrate the point that a state testing program need not be limited to knowledge and basic skills. Here is a readily available instrument that can be used to measure interests in school subjects in a convincing manner. It is a list of 192 activities, 16 in each field of study, like "Talk about books in class," "Compare accounts of the same event in different newspapers," "Collect and classify plants," and "Typewrite business letters." They are listed in seemingly random order, and the directions amount to marking each activity 2 if you like it or think you would like it, 1 if you don't know

or don't care one way or the other, or 0 if you dislike it or think you would dislike it. The score for each field is the sum of these numbers, and the two columns headed Mean show the averages of a nationally representative sample of 15,450 juniors in 187 high schools.

Note how widely the academic interests of boys and girls differ. Industrial Arts, for example, was highest for boys, lowest for girls. The rank-difference correlation between these two columns is -.70. Boys placed Physical Sciences, Biology, and Mathematics near the top, girls near the bottom. Girls liked Foreign Languages, Art, and English far better than boys. Both sexes preferred Foreign Languages to English by two ranks and Art to Music by two and four ranks respectively. Both gave high rank to subjects with vocational possibilities: Industrial Arts and Business for boys, Home Economics and Secretarial for girls.

If you wonder where Physical Education stood, the answer is that it was dropped from the instrument because it was such an overwhelming favorite with both sexes in every school that we learned nothing by including it. Its omission allowed us to break up the large field of Business Education into two parts, labeled somewhat ambiguously as Business and Secretarial. They refer to the offerings most commonly elected by boys and by girls. Although there is some overlap in the activities representing these fields, their separation permitted Business to rank third for boys and Secretarial second for girls, where formerly the combined field of Business Education ranked about fifth for both sexes.

The advantage of using this instrument (called AIM for "Academic Interest Measures") is that it yields a precise numerical score for each field, which gives teachers a realistic target to shoot at and try to change. As English teachers, we are naturally concerned that English ranks third from the bottom for boys and two ranks below Foreign Languages for both sexes. If we tried to

move it up to the top, we would be pretty sure to be disappointed, because fields like Industrial Arts for boys and Home Economics for girls have a natural appeal with which it is hard to compete. But certainly we can do better than that average score of 13.51 for boys, and the beauty of this instrument is that if we got it up even as far as 15, we and our principal and counselors would know that we were making progress. Without the comparative information provided by a large study, there would be no way to judge whether an average of 15 was good or bad.

My fourth exhibit is a table entitled "Percent of Students with the Following Scores on the Newspaper Test." I included it to show that once in a while in large-scale testing one can prove that some program is a glorious and heart-warming success. That is not the kind of language in which test results are generally reported. My fellow testers are so conservative that I sometimes think the only conclusion they ever reach is that more testing is needed. But these figures seem to me to constitute proof that the "Newspaper in the Classroom" program, generously supported over the past 14 years by the American Newspaper Publishers Association, has made a substantial difference in students' understanding of newspapers.

The test was based on two simulated newspapers that students read and referred to while they took the test, and there were 30 four-choice questions in each form. Scores were compiled for two types of classes: newspaper classes that had been involved in the "Newspaper in the Classroom" program, and regular classes that had made no systematic use of daily newspapers in school. The table compares results at four score levels, from scores of 10 or better up to scores of 25 or better, and you can see that the newspaper classes were consistently and substantially superior in both junior and senior high schools. These results were based on about 13,000 students in three large areas in different regions of the country. The decision to test

in these three areas was made last March; the test was given in May; and most teachers heard about it only a few days before the test. While their programs were going on, neither the students nor their teachers knew that they would be tested at the end, and they had no conception of what the test would be like. These areas had not been involved in the development of the test, and it had not been used previously in any of these schools.

There were many attempts to explain away these differences. "Aren't these newspaper programs generally found in the better schools of prosperous neighborhoods?" Possibly, but in this study the newspaper and regular classes were drawn from the same schools in the same neighborhoods. "Aren't the newspaper teachers likely to be superior?" We like to think so, but nearly every teacher of a newspaper class also contributed one or more regular classes; they do not teach the newspaper in all their classes. "Aren't the newspaper students likely to be better readers to begin with?" Why would any principal go to the trouble of putting his superior readers into newspaper classes? He would assume that they, if anyone, could read newspapers adequately. If there is any difference of this sort, it is usually the less capable readers--the disadvantaged--who are given newspapers to engage their interest--to show them that reading has some connection with the real world. "But won't any group do better if it is given special treatment, as in the well-known Hawthorne Effect?" Yes, but no one regards this program as experimental. It has been going on for fourteen years--not quite that long in these three areas, but they have some of the most solidly established of all newspaper programs. It is a truism among researchers that almost anything will work once, but after the novelty wears off, differences tend to fizzle out--especially in a field like reading, which is more closely linked to maturation than anything else we teach. It is almost as hard to add a cubit to one's stature by taking thought as to improve any kind of reading in high school beyond the levels ordinarily attained. Yet here is a program concentrated on

the reading of newspapers--the most frequent and possibly the most important reading of the average citizen--that still produces not only significant but also socially important differences in students' understanding of newspapers after fourteen years. My colleagues in research at ETS were unable to recall anything like it among all the reading improvement programs we have evaluated. I find it hard to believe that it was superior teaching that made the difference. My guess is that it was just regular exposure to good newspapers, aided by students' arguments with one another over what the articles meant.

What about the performance of regular students who had never been involved in this program? In junior high, their average score was slightly less than half right; in senior high, slightly less than 60 percent right. The questions were not designed to be hard or tricky. They were written and reviewed by committees of teachers with the idea that, if a student understood an article, he should be able to answer the questions. Remember that every student had a copy of the simulated newspaper to read and refer to while he was taking the test. There was no pressure of time; in this administration all students had a full class period to answer 30 questions, and there were very few incomplete papers. If, under these conditions, the average student gets only 50 to 60 percent of the questions right, his understanding of newspapers is seriously deficient.

As a matter of curiosity, I looked into the background of the regular student in junior high school who stood nearest the national average. He was 13 years old, in the eighth grade, and had been in the same school since the fourth grade. He was in the 120 to 125 IQ range and was making B's and a few A's in his subjects. He was especially good in math and science, played chess, and liked to read history and science fiction. His principal estimated that the family income was approximately $20,000 with the father employed as the labor relations representative of a nationally known manufacturer and

the mother at home. They live in an upper middle income residential area, adjacent to high income homes.

There was nothing "disadvantaged" about this student. He might be your son or mine, and we would be proud of him. Yet he could answer only half the questions correctly about what typical newspaper articles said. If he stays at the national average, in senior high he will be able to answer about 60 percent of such questions correctly. At age 18 he will vote.

Something ought to be done about it, and the results of this norming study seem to me to constitute proof that the "Newspaper in the Classroom" program is something that works.

I have now illustrated four kinds of things that statewide testing can do. It can bring to light the superior results of a group of schools where no one would expect it and raise interesting questions about how they managed it. If it deals with the generally lower scores of disadvantaged minorities, it can put these differences in perspective by showing comparable differences between boys and girls. It can deal with school effects other than knowledge and basic skills, as shown not only by the interest measure but also by the data on attitudes toward school in the first exhibit. Finally, it can occasionally show that a particular program like the newspaper program is producing substantial and socially important results. Three of my exhibits were based on areas larger than a state, but it is obvious that the same kinds of information could be secured in any state testing program.

Two of my exhibits also illustrated possible dangers in state testing programs. In the first, we noted that the small towns in a poor region that did exceptionally well were in danger of losing state aid as a consequence of the generally sound policy of using such aid to overcome deficiencies. But remember that all the information needed to bring about a salutary revision of this policy was contained in the reports of results. Without a state testing program, no one would have guessed how well these towns were doing.

The second exhibit might have led to conclusions prejudicial to Negroes if only one graph had been presented, but the saving grace of a large study is that there is always more than one graph. The comparable positions of boys and girls put these differences in proper perspective.

As in any human enterprise, a state testing program offers at least as many possibilities of doing harm as of doing good. The only way I know to insure a preponderance of good is eternal vigilance.

I should like to add one suggestion for your state testing program that has nothing to do with my exhibits. I see no point in giving the same tests to everybody in the same grade in the whole state if your main concern is to discover the strong and weak points in your educational program. I would insist on getting up at least ten different test packages (with just one short test, probably vocabulary, common to all packages for equating purposes), and I would arrange these packages in such fashion that the first student in each class tested would get Package 1, the next Package 2, and so on around the class. As it is now, with the usual limitations of testing time, you may have to cover all teaching of literature in something like 30 minutes. With ten packages, that would be extended to 300 minutes--a total of five hours of testing time--without taking more time from any given student or class, and you would still have an adequate random sample of students taking each package. Obviously you can find out a great deal more about the teaching of literature in 300 minutes than you can in 30 minutes, and the more extensive sample is less likely to influence the teaching of literature to its detriment.

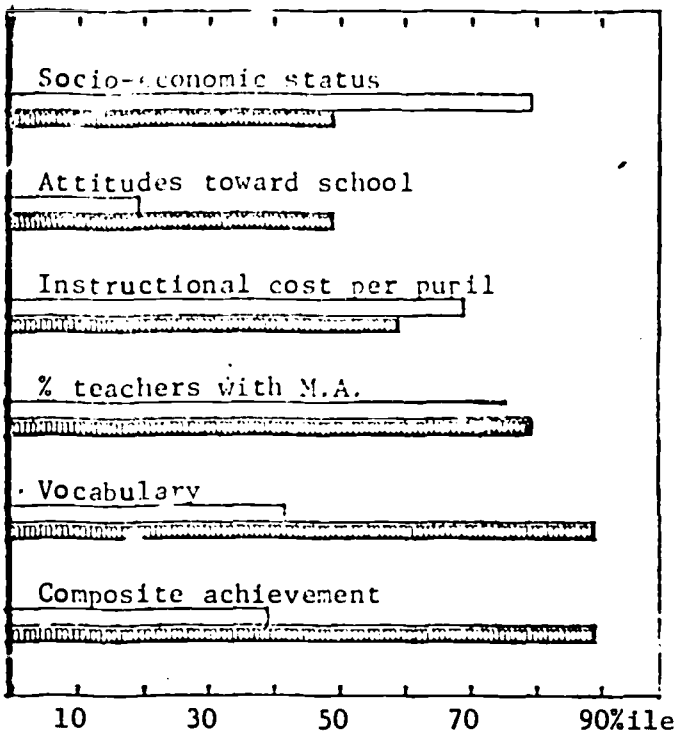
Think about it. It is entirely feasible; it is the way the National Assessment operates; and I am sure that the same salutary principle can be extended to state testing programs with beneficial results.

November 11, 1971

Prepared to talk by Paul B. Diederich, ETS, in Saturday (Nov. 27) [illegible]
Discussion Group "State-Wide Testing - Two Views."

Paul B. Diederich, EG-16
Educational Testing Service
Princeton, N.J. 08540

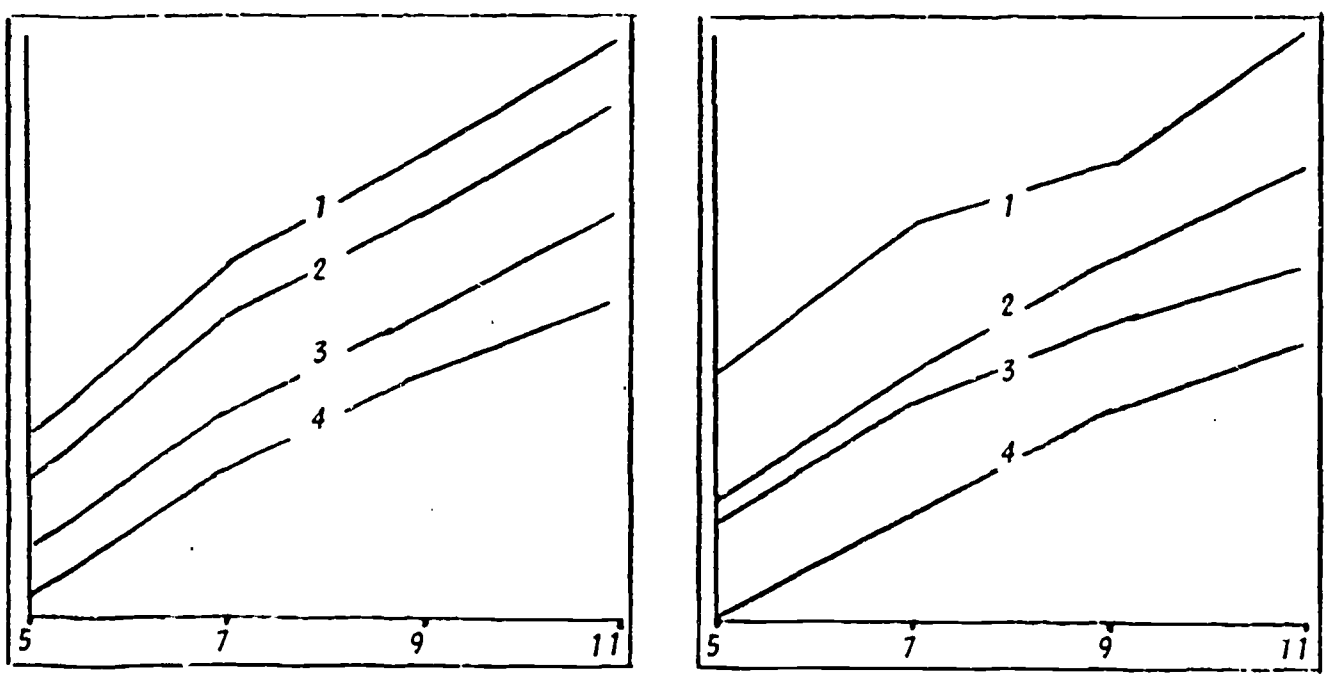
Selected Indices for Towns in Two Regions of One State



Grade 4

Averages for towns of 2,500 to 10,000 population are plotted on the distribution of averages of all 600 districts in the state. Although "white" towns have more money, "shaded" towns are 30 percentiles more favorable in attitudes toward school. In spite of lower costs per pupil, they also have slightly more teachers with M.A. degrees. The big surprise is that "shaded" towns stand at the 90th percentile in vocabulary (a rough index of brightness) and in composite achievement (reading, writing, arithmetic); "white" towns at the 40th. These differences are unrelated to race; neither region has a minority problem, and "white" towns are wealthier. How did the "shaded" towns manage it?

Are Boys Genetically Inferior to Girls? or Negroes to Whites?



Average Scale Scores on STEP Reading in Grades 5, 7, 9, and 11

1 Girls Academic	3 Girls Nonacademic	1 White Academic	3 White Nonacademic
2 Boys Academic	4 Boys Nonacademic	2 Negro Academic	4 Negro Nonacademic

OVER

Rank Order of Interests in School Subjects Revealed by AIM, 1970

Boys' Interests			Girls' Interests		
Rank	Field	Mean	Rank	Field	Mean
1	Industrial Arts	22.55	1	Home Economics	25.35
2	Physical Sciences	20.01	2	Secretarial	22.39
3	Business	18.22	3	Foreign Languages	20.79
4	Biology	17.39	4	Art	19.76
5	Social Studies	17.34	5	English	18.75
6	Mathematics	17.09	6	Business	18.71
7	Secretarial	16.02	7	Social Studies	17.03
8	Foreign Languages	14.99	8	Music	16.57
9	Art	14.83	9	Biology	15.68
10	English	13.51	10	Mathematics	12.86
11	Music	13.45	11	Physical Sciences	11.70
12	Home Economics	12.61	12	Industrial Arts	10.96

AIM = Academic Interest Measures, published by ETS. The figures above are based on a representative national sample of 15,450 juniors in 187 high schools. Note that boys and girls differed widely in their expressed interests: e.g. Industrial Arts was highest for boys, lowest for girls. (The rank-difference correlation between these two columns is -.70.) Boys placed Physical Sciences, Biology, and Mathematics near the top, girls near the bottom. Girls liked Foreign Languages, Art, and English far better than boys. Both sexes preferred Foreign Languages to English by two ranks, and Art to Music by two and four ranks respectively. Both gave high rank to subjects with vocational possibilities: Industrial Arts and Business for boys; Home Economics and Secretarial for girls.

Percent of Students with the Following Scores on the Newspaper Test

	Junior High School		Senior High School	
Score	Regular Classes	Newspaper Classes	Regular Classes	Newspaper Classes
10 or better	79%	87%	86%	94%
15 or better	47%	60%	65%	78%
20 or better	17%	28%	36%	51%
25 or better	2%	4%	11%	17%

In May 1971 the ANPA Foundation Newspaper Test was given to 13,000 students--9,000 in junior high and 4,000 in senior high--in and around Charlotte, Peoria, and Fort Worth, representing middle-size cities served by good newspapers with active "Newspaper in the Classroom" programs. "Regular classes" were those that had made no systematic use of daily newspapers in connection with school work; "newspaper classes" had done so and were usually involved in a "Newspaper in the Classroom" program. The test was based on simulated newspapers of four pages each which were read and referred to during the test. There were 30 four-choice questions in each form of the test. The table shows the consistent superiority of newspaper classes.