DOCUMENT RESUME

ED 171 820 UD 019 333

AUTHOR Astin, Alexander W.
TITLE Testing in the Post "Bakke" Period.
PUB DATE Apr 79
NOTE 15p.; Address presented at the Annual Meeting of the American Educational Research Association (San Francisco, California, April 11, 1979); Not available in hard copy due to the reproduction quality of the original document

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Academic Achievement; Admission Criteria; Affirmative Action; *College Admission; *Equal Education; *Grade Prediction; Higher Education; *Minority Groups; Norm Referenced Tests; *Standardized Tests

ABSTRACT

Efforts to implement affirmative action programs will not be effective until alternatives to the traditional use of standardized tests in admissions, grading, and tracking are considered. These test scores are used in college admissions to predict college grades. Some kind of selective admissions process is needed, but admissions procedures should be designed to select students, and not to track them. Modifying admissions policies is the best way to remedy the unequal opportunity posed by tracking. Alternatives to the standardized test for admissions must be found if minorities are to gain equal access to higher education. (Author/MC)

TESTING IN THE POST "BAKKE" PERIOD*

Alexander W. Astin
University of California, Los Angeles
and
Higher Education Research Institute

One of the most gratifying outcomes of the controversy surrounding the recent Bakke case was that the higher education establishment lined up squarely behind the University of California supporting minority-oriented affirmative action plans in college admissions. Even before the decision came down, representatives of the major educational associations expressed their determination to continue and even expand minority-admissions programs.

While these public pronouncements are surely reassuring to those of us who are committed to expanding educational opportunities for minorities, it is not clear just how such lofty intentions are to be translated into concrete actions that will, in fact, strengthen affirmative action efforts. No matter how committed they may be to minorities in theory, most educators are reluctant to recognize that some of their time-honored practices still pose serious obstacles to equal access for minorities and the disadvantaged. Specifically, the activities I have in mind are admissions, grading, and ability tracking. Standardized tests, it should be noted, play a prominent part in each of these activities. The thesis of my talk today is that our efforts to implement affirmative action programs will not be very effective until we are willing to consider alternatives to the traditional use of standardized tests in admissions, grading, and tracking.

Most of the controversy about standardized testing has emphasized

*Address presented at the 1979 Annual Meeting of the American Educational Research Association, San Francisco, April 11, 1979.

what might be called the "construct validity" of these instruments. Are they "culturally biased"? Do they accurately reflect the academic abilities of persons who are not from white, middle-class backgrounds? In my judgment, these concerns have been misplaced, since the controversy over what the tests "really" measure has tended to obscure a much more fundamental problem: The way the tests are used is basically inappropriate to the purposes of education. This misuse of tests, in turn, poses special obstacles to the educational development of disadvantaged minority groups.

Before considering my specific objections to current uses of tests in admissions, grading, and tracking, it should be made clear that my arguments are predicated on one fundamental assumption: namely, that the major purpose of higher educational institutions is to educate students. In other words, a fundamental mission of any college or university is to produce certain desirable changes in the student or, more simply, to make a difference in the student's life. This conception of the purpose of higher institutions somewhat resembles the economist's notion of "value added."

Let's first consider the use of tests in admissions. Perhaps the best way to understand the use of tests in admissions is to ask an admissions officer or faculty member why tests are used to decide which students are should be admitted or rejected. The most likely response to this question is, "because the tests predict grades in college." Forgetting for the moment that the college grade point average leaves something to be desired as a measure of student learning, let us simply consider the educational implications of the prediction argument: People with high scores on an admissions test are preferred over those with low scores on the grounds that they will subsequently earn higher grades in college. In my judgment this prediction argument is a weak justification for the use of any selection device, simply

because prediction may have little, if anything, to do with the educational or "value added" mission of the institution.

To illustrate the fallacy inherent in the prediction argument, suppose that a college was admitting all the wrong students, so that the students learned absolutely nothing from their college experience. Even in such an extreme case, the admissions tests could still predict grades. In the same vein, if one could administer college admissions tests to high school seniors, put them in a state of suspended animation for four years, then revive them and give them a set of final examinations, the college admissions tests would still have "validity" in predicting performance on the final examinations. As a matter of fact, since different students are likely to learn at different rates, it stands to reason that the greatest student learning may be occurring at those colleges where tests have the poorest predictive "validity."

A possible rejoinder to this objection is that the criterion used to justify admissions testing--the college grade point average--is indeed a measure of what the student has learned in college. While there has been very little research on this fundamental problem, what evidence there is fails to support it. One study (Harris, 1970) examined pretest-posttest gains on the College Level Examination Program administered before and after exposure to various courses. As it turns out, students who got failing or near-failing grades showed test score gains comparable to those of students with high grades. Is it possible that standardized testing, when carried out on a pretest-posttest basis, offers a much better means of assessing student progress than the grade point average? Should our dependence on tests shift from its current emphasis on selection and prediction to a greater emphasis on learning and development?

The main objective to traditional grading practices, of course, is

that the college grade point average is a relative measure that simply ranks students from best to worst without any necessary reference to what the student has learned as the result of the educational program. Why institutions have not relied more on standardized tests to assess cognitive learning has always been a puzzle. Many professors argue that the tests are somehow inadequate: too superficial, incomplete in coverage, culturally biased, and so forth. Professors' heavy reliance on these same tests to decide undergraduate admissions, advanced placement, and admission to graduate and professional schools belies these objections. If academics are perfectly willing to use standardized tests to evaluate the performance of prospective students, how come such tests suddenly become inappropriate measures of learning once the student is admitted?

One difficulty in using standardied tests to monitor student development is the way they are scored. Virtually every test-maker today reports student performance only in normative terms: percentiles, standard scores, stanines, and so on. These norm-based scores show how the individual performs only in relation to others. Such scores are useful for selecting and screening (i.e., identifying the "best" students), but they are difficult to use for measuring change or growth. Lacking any absolute referent, it is difficult to know from successive administrations of such tests if, or how much, the person's performance level has improved. (The traditional letter grade system suffers, of course, from the same limitation, particularly if students are graded on the curve.)

There is no good reason why such tests cannot be used to produce information on changes in student's absolute level of performance: for example, in the number or percentage of items answered correctly. In addition, performance data on individual test items could be provided to diagnose

particular strengths and weaknesses or to measure growth in more specific areas of skill or knowledge. Psychometricians have discouraged testing companies from reporting results for individual test items on the grounds that such information is unreliable. However, individual item data can be highly reliable if they are reported for groups of students.[1] Many colleges and universities that use standardized tests demand that results be reported in raw score form and that item data be included, the testing companies will probably provide them. Such demands could be made for most of the tests now in widespread use: the achievement and advanced placement tests of the College Entrance Examination Board (CEEB), the College Level Examination Program tests, the undergraduate program of the Graduate Record Examination, and the various devices to select students for graduate and professional schools. Considering the additional information that would be available to teachers and students, the cost of providing it would be trivial. Given that so much care is devoted to writing and pretesting items and that these activities account for a major share of the cost of standardized test construction, the potential benefits from these developmental efforts should not be lost in the construction of "scales" and the computation of percentiles, standard scores, and other normative measures.

Why do the testing companies persist in limiting their feedback to normative measures when such information is of limited value to students and institutions? There seem to be at least two explanations. First, the psychometricians who control the technical aspects of the test industry have become mesmerized by the statistical properties of standard scores and by the elegance of the normal distribution which underlies classical test theory. Thus,

1. Data on individual test item performance has been used, with considerable success, in the International Study of Achievement in Mathematics, edited by T. Husen (New York: Wiley, 1967). (Stockhom, Almquist and Wiksell.)

raw scores (number answered correctly, rights minus wrongs) are converted to normative scores, a procedure that eliminates the original units of measurement. From the purist's perspective, individual items have even worse statistical properties, so they must be lumped together in sufficient numbers to produce "scales" that form the appropriately shaped distribution with the proper degree of "reliability."

Second, and more important, is that normative measures are consistent with the competitive and meritocratic values that permeate so much of American society. In this regard, it should be kept in mind that the first individually administered intelligence test, developed early in this century, served as a model for the group-administered tests used today. While the score was not, strictly speaking, a normative measure, it prompted the use of normative jargon with a strong meritocratic connotation: genius, superior, dull normal, imbecile, and so forth. The earliest large-scale applications of group testing were in the military, which, during the two world wars, was concerned with screening out illiterate and "mentally defective" recruits and draftees and identifying candidates for officer training. These applications were basically meritocratic: finding the "best" and "worst" candidates. This view of group tests continued after World War II when the crush of applicants forced many colleges to institute screening procedures that could be applied on a large scale at relatively low cost. Normative scores provided a simple and seemingly fair means to identify the "best" students. This meritocratic view of testing has been reinforced by the competitveness of the colleges themselves, where the "best" colleges are, of course, those with the highest scoring students.

In the 1950s and 1960s this meritocratic orientation was reinforced by competitiveness at the international level, when many Americans interpreted the first Soviet Sputnik to mean that the United States had slipped behind

because many of its "brightest" students were not going to college. One manifestation of this concern was the National Merit Scholarship Corporation, which annually came close to one million students just to identify the 1,500 or so with the highest scores who could be awarded scholarships to assure college attendance. Colleges, of course, became highly competitive in their quest for Merit Scholars, and the number of Scholars in the student body was widely regarded as a sign of academic quality. A similar competition developed among high schools.

Today, most of us in academe take the normative-meritocratic nature of testing for granted. As a consequence, concern with student change, growth, or development is subordinated to ranking students from best to worst. This bias has infected grading practices, where instead of determining what students learn from the beginning to the end of a course, or what they learn in college, grades are assigned which reflect primarily how students perform compared with each other at a given point in time. Employers and graduate schools further reinforce these grading practices, since the student's GPA is a convenient means to identify the "best" students.

There is no reason why colleges should persist in these practices other than habit and tradition. Using results from individual items and absolute or raw scores requires little additional effort, regardless of whether the test is a nationally standardized or individual classroom examination.

It is important to recognize that a meritocratic approach to selection of talent probably makes much better sense in other sectors of our society such as the military, business and industry. What seems to have happened is that college admissions officers now function more like personnel managers in a commercial enterprise than like educators. Picking the "best"

is an appropriate activity for business and industry, since their goals are to hire the best talent to maximize productivity and profit. Similarly, competition among rival companies for the pool of available talent is consistent with the very nature of free enterprise economy. But the business model, which has been adopted by the most selective institutions, is not appropriate to education. The college does not exist primarily to exploit talent. The mission of the college is not simply to maximize its output of distinguished alumni by enrolling as many talented students as possible. Such a static process reduces the college to a kind of funnel: What comes out is purely a matter of what goes in. Colleges and other educational institutions exist to develop talent, to change the student, to contribute to personal development, to make a difference.

If colleges and graduate and professional schools must limit the number of available spaces for applicants, then some kind of selection process is obviously needed. The "value added" approach to education suggests that admissions procedures should be designed, at least in part, to select students who are likely to be influenced by the educational process. The difference between the value added and the predictive approaches can be illustrated with an analogy from the field of thoroughbred racing. Some years ago I put forward the argument that admissions officers in selective institutions function very much like handicappers at the race track: They try merely to pick winners. By looking over the various candidates and evaluating their respective talents, they select those likely to perform well. But handicappers are interested only in predicting the horse's performance, not in helping it to run better and faster. An educational institution should function not like a handicapper, but like a jockey or trainer: It is responsible for improving the performance of the student, not merely

for identifying those with the greatest potential.

My argument so far involves two basic proposals for change: (a) first is that the meritocratic model of admissions be abandoned in favor of a value added approach and (b) that test scores, together with item and raw score data, be used instead to provide both students and teachers with feedback concerning student growth and development. While this second proposal could probably be implemented tomorrow if the academic community were prepared to abandon its traditional grading system, the "value added" approach in the admissions process poses a number of unresolved problems. Most important is that there are probably no existing selection devices that adequately reflect the student's potential to benefit from the collegiate experience. Lacking such measures, it might be argued that we should stick to traditional admissions tests until something better is found. The problem with this argument is that the negative effects of tests in college admissions are not benign: they operate to the particular disadvantage of those minority groups who represent some of the most disadvantaged segments of our society.

We recently conducted a simulation survey to determine just how serious the handicap posed by college admissions tests really is. For example, if we are faced with a crush of applications where only one in ten of the applicants can be admitted (such a situation is not unlike that found in many medical and law schools), relying solely on tests as our selection guide eliminates 99 percent of the black applicants and 97 percent of the Chicano applicants. In other words, a white student is three times more likely to be selected than a Chicano student, and ten times more likely than a black student. Clearly, some other approach to the admissions process should be found if these minority groups are ever to reach anything resembling educational parity with the white majority.

Let us now turn to the third educational practice that involves the use of standardized testing: tracking of students into different types of institutions. Public higher education in many states is arranged hierarchically, with the major public university occupying the top rung, the state colleges (many formerly teachers colleges) the middle rung, and the bottom position being occupied by open-door community colleges. In some states, tests play a major role in determining which of these three options is available to the high school graduate. Generally speaking, the highest-scoring students can avail themselves of all three tiers, whereas the lowest-scoring students are allowed to enter only the community colleges. Not surprisingly, low income and minority students tend to be underrepresented in the universities at the top of the hierarchy and heavily concentrated in the community colleges.

Now if these different types of institutions were roughly equivalent in their resources and offerings, one could argue that this type of ability tracking is not a denial of equal access. However, institutions at different positions in the hierarchy are by no means equivalent, so that the student's future may depend as much on the kind of institution attended as on attendance versus nonattendance. With the proliferation of public community colleges and the substantial financial aid now available to needy students, the real issue of equality of access is not who goes to college, but who goes to college where.

If one compares the educational resources of these three levels of institutions, in every respect they form a perfect hierarchy with the universities having the greatest resources and the community colleges the least. I am speaking here of such varied attributes as student-faculty ratios, faculty salaries, physical plant, libraries, and expenditures for general educational

purposes. Moreover, institutions at the top of the hierarchy are also much more likely to have another asset: residential facilitiies. A considerable body of recent research (Astin, 1977: Chickering, 197) shows that students get more out of their undergraduate experience if they can live on campus rather than commute from home. Finally, longitudinal studies show that a student's chance of persisting is substantially better at a four-year college or university than at a community college. In short, these results suggest that the use of tests to track students into different types of institutions substantially reduces educational opportunities for those students who tend to receive low scores.

Is there any educational justification for the tracking arrangement that exists in most states? Perhaps the most common rationale offered for tracking is that students will develop better academically if they are grouped with students of similar ability. In effect, this argument assumes a kind of aptitude-treatment interaction effect which has several important corrolary assumptions. First, high scoring students are [illegible] to need the stimulation and competition of each other to realize their full potential. Second, it is assumed that high scoring students will become bored and apathetic if grouped with lower-scoring students. And finally, it is argued that the lower-scoring student will become intimidated and discouraged if forced to compete with high scoring students. Considering the fundamental importance of such interaction assumptions, it is remarkable that so little research has been done to test them. What evidence there is, however, offers virtually no support. Thus, as far as learning outcomes are concerned, there seems to be little or no interaction between the selectivity of the institution and the ability of the student (Astin, 1968: Nichols, 1964; Rock, Centra, and Linn, 1970).

My hunch is that institutional tracking is perpetuated less for educational reasons than for reasons of meritocratic competition and status. Professors in the universities at the top of the hierarchy support selective admissions because students with high test scores are easier to identify with and easier to teach. Indeed, even within a given classroom, professors probably favor their most advanced students. Selective admissions is supported by alumni, legislators, faculty, administrators, and probably a great many students because a good input of highly motivated and talented students will almost guarantee a good reputation and a good output of distinguished and possibly wealthy alumni. The secondary schools support the track system that results from selective admissions because they see it as a reward or incentive system to motivate their students: Teachers and guidance counselors frequently tell their students to study hard so they can get into a "good" institution.

Selective admissions is also justified in these institutions on the grounds that any relaxation of admissions standards at the selective institutions would lower their academic standards. While such an outcome is indeed possible, it is by no means inevitable. Academic standards have to do with the performance levels required before the institution will certify that the student has passed certain courses or completed certain requirements for the degree. There is no necessary reason why a modification in admissions standards has to be accompanied by a modification in certification standards.

Perhaps the most important reason for hierarchical public systems is economic. In the minds of many legislators and planners, the two-year college is an appealing way to expand access because it is much less expensive than other institutional forms. At the same time, expanding the community college allows the more selective universities to preserve their selectivity

and prestige and to avoid the pedagogical difficulties associated with teaching less-well prepared students. Viewed from the perspective of the minorities, however, these hierarchical public systems based on selective admissions represent a denial of equal opportunity.

There would seem to be at least two ways to approach the problem of how to remedy the unequal opportunity posed by hierarchical public systems. The first, and certainly the most unrealistic approach, would be to equalize resources and expenditures across the different types of institutions. If this means taking resources away from the universities to upgrade the community colleges, the universities would never put up with it. If it means adding resources to the total higher education budget within the state, the legislators and taxpayers would probably never support it. The second approach--which involves modifying admissions policies--seems like a much simpler and certainly less expensive solution. Nevertheless, if the universities change their policies to admit larger numbers of underprepared students, additional resources will almost certainly be required to provide adequate remedial assistance and other special support services.

In conclusion, I suppose I find myself on two sides of the testing issue: advocating less dependence on tests for one purpose and more use of tests for another. Specifically, I think the time has come to deemphasize the use of standardized tests in screening and selection and to reconsider their potential for evaluating student progress and assessing the effectiveness of educational programs at all levels. In research jargon, this means that tests should be used less as independent variables and more as dependent variables. At the same time, we should begin to consider retrieving some of the information we lose when we aggregate test items into scales and convert raw scores into normative scores. My impression is that the resistence to

standardized testing will decline substantially if people can come to regard these devices as sources of feedback to enhance the teaching-learning process, rather than as screening devices that limit educational opportunity.

But what about the admissions process? If the use of tests is to be minimized in the interest of expanding opportunities for minorities, and if the predictive model is to be abandoned in favor of a "value added" model, how are institutions that have more applicants than available spaces supposed to make their selection decisions? While I have no magical solutions for this problem, selective institutions in the public sector might want to consider experimenting with alternative selection procedures identifying students with the greatest potential for growth or learning. The ideal solution, of course, would be to reshape our public systems to the point where there is a much closer correspondence between student demand and the supply of available places in various types of institutions. Such a system, of course, would obviate the need for selective admissions. But until such a utopian balance is achieved, the standardized test and the predictive model of admissions will continue to represent serious barriers to the educational development of minority students. If we are really serious in our intentions to expand access for minorities, alternatives to the standardized test must be found.