DOCUMENT RESUME

ED 155 205 TM 007 155

TITLE	Educational Measurement & the Law. Proceedings of the 1977 ETS Invitational Conference.
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	78
NOTE	113p.; Proceedings of the Educational Testing Service Invitational Conference (38th, New York, New York, October 29, 1977)
AVAILABLE FROM	Invitational Conference, Educational Testing Service, Princeton, New Jersey 08541 ($5.00)
EDRS PRICE	MF-$0.83 HC-$6.01 Plus Postage.
DESCRIPTORS	*Admission Criteria; Awards; Cheating; College Admission; College Entrance Examinations; *Conference Reports; *Constitutional Law; Court Litigation; Decision Making; Disadvantaged Groups; Due Process; Educational Legislation; *Educational Testing; Equal Education; Equal Opportunities (Jobs); *Equal Protection; Evaluation; Evaluation Needs; Federal Legislation; Graduate Study; Higher Education; *Legal Problems; Minority Groups; Occupational Tests; Predictive Validity; Professional Education; Reverse Discrimination; Screening Tests; Test Bias; Testing Problems; Test Interpretation; Test Validity
IDENTIFIERS	Bakke vs Regents of the University of California; De Funis v Odegaard; Elementary Secondary Education Act Title I; Griggs v Duke Power Company; Testing Industry; Washington v Davis

ABSTRACT

At the 1977 Educational Testing Service (ETS) Invitational Conference, the ETS Measurement Award was presented to Anne Anastasi. In view of the convergence of measurement and the law, the conference focused on six related issues. Barbara Lerner explored the screening procedures of American professional and graduate schools in "Equal Protection and External Screening: Davis, De Funis, and Bakke," and was responded to by Ernest M. Bernal, Jr., and Deane C. Siemer. Melvin R. Novick discussed funding allocations under Title I, admissions policy and the Bakke case, federal guidelines for employment testing, and due process in the handling of suspected cheating cases, in "The Influence of the Law on Professional Measurement Standards." Winton H. Manning responded to Novick's presentation. In the final session, Wayne H. Holtzman reviewed the implications of several court cases in "Validity and Legality;" Charles L. Thomas discussed "Some Possible Social Implications of Recent Court Decisions;" Norman Frederiksen offered some ideas for improvement in test use in "There Ought to Be a Law;" and Michael Scriven suggested what measurement experts might learn about decision making from the law, in "The Logic of Judgment in Evaluation and the Law: Making Hard Decisions with Soft Data." (BW)

PROCEEDINGS OF THE 1977 ETS INVITATIONAL CONFERENCE

Educational Measurement & The Law

EDUCATIONAL TESTING SERVICE

Educational Measurement & The Law

PROCEEDINGS OF THE 1977 ETS INVITATIONAL CONFERENCE

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY

ATLANTA, GEORGIA
AUSTIN, TEXAS
BERKELEY, CALIFORNIA
EVANSTON, ILLINOIS
LOS ANGELES, CALIFORNIA
SAN JUAN, PUERTO RICO
WASHINGTON, D.C.
WELLESLEY HILLS, MASSACHUSETTS

The thirty-eighth ETS Invitational Conference, sponsored by Educational Testing Service, was held at the New York Hilton, New York City, on October 29, 1977.

Presiding: William W. Turnbull
President
Educational Testing Service

ETS Invitational Conference Program Committee

Chairman: Scarvia B. Anderson
Jack R. Childress
Winton H. Manning
Samuel J. Messick
Janice Somerville
Warren W. Willingham
Jane D. Wirsig

Conference Coordinator: James R. Deneen
Assistant Coordinator: Margaret B. Lamb
Editor, Conference Proceedings: Nathaniel H. Hartshorne

Library of Congress Catalog Number: 47-11220
Printed in the United States of America.

Contents

v Foreword by William W. Turnbull

vii Presentation of the 1977 Measurement Award to Anne Anastasi

Session I

3 Equal Protection and External Screening: Davis, De Funis, and Bakke
Barbara Lerner

29 Discussion: Ernest M. Bernal Jr.

35 Discussion: Deanne C. Siemer

Session II

41 The Influence of the Law on Professional Measurement Standards
Melvin R. Novick

53 Discussion: Winton H. Manning

Session III

63 Validity and Legality
Wayne H. Holtzman

73 Some Possible Social Implications of Recent Court Decisions
Charles L. Thomas

87 There Ought to Be a Law
Norman Frederiksen

99 The Logic of Judgment in Evaluation and the Law: Making Hard Decisions with Soft Data
Michael Scriven

Foreword

Educational Measurement and the Law, the theme of the 1977 ETS Invitational Conference, was timely for several reasons. Public interest in the *Bakke* case is an insistent reminder of a new fact: that what people do in the field of measurement or what uses they make of the results of measurement are no longer the concern only of the testing fraternity or indeed of educators or employers. These actions have entered the realm of public concern and debate.

In view of this convergence of measurement and the law, the 1977 Invitational Conference was intended to bring together people with expertise in both these areas to discuss some of the critical issues. The conference focused on six aspects of measurement and the law in three sessions. Barbara Lerner, the first speaker of the morning, explored the screening procedures of American professional and graduate schools and the conflicts that have arisen as a result of our efforts to maintain the rights of both individuals and of groups. In the second session, Melvin Novick took a close look at five critical areas of educational measurement and the ways in which they are affected by the law and suggested a number of actions that might be taken to better deal with those situations.

Two of the afternoon speakers were concerned with technical and social aspects of the theme. Wayne Holtzman, in his paper on "Validity and Legality," reviewed a number of court cases and discussed their concern with several types of test validity. Charles Thomas focused on some of the reasons for legislation designed to increase opportunities for minorities and the concerns expressed by proponents and opponents of such affirmative action programs. In his paper, *There Ought to Be a Law*, Norman Frederiksen offered some clear-headed and practical ideas for change and improvement in the use of tests and testing. Michael Scriven, the final speaker of the day, suggested what the field of measurement might learn from the law about the complex processes of decision making.

It was a rich day with stimulating presentations for which we are most grateful not only to the speakers mentioned above but also to Ernest Bernal, Deanne Siemer, Thaddeus Holt, and Winton Manning whose discussions of the morning presentations added an important critical dimension to the conference, and to Roger Lennon and Alfred Fitt who presided over the second and third morning sessions.

Looking back at the topics discussed at the 1977 Invitational Conference, one is reminded again of the truth of E. B. White's prediction of a bright future for complexity. It is perhaps a mark of our maturity as a field that we are beginning to confront that complexity more squarely.

William W. Turnbull
PRESIDENT
Educational Testing Service

EDUCATIONAL TESTING SERVICE

Measurement Award 1977

ANNE ANASTASI

Anne Anastasi

The ETS Award for Distinguished Service to Measurement was established in 1970 to be presented annually to an individual whose work and career have had a major impact on developments in educational and psychological measurement. The 1977 Award was presented at the ETS Invitational Conference by President William W. Turnbull to Dr. Anne Anastasi with the following citation:

Recognized internationally as a leading exponent of differential psychology, Anne Anastasi has exerted a major influence on its development as a behavioral science. Educational measurement has benefited significantly from her inquiries into the nature of individual differences.

She has confronted the controversy regarding the effects of heredity and environment on abilities, clarifying the joint role of these conditions in human development, and sweeping away many of the presuppositions, myths, and fallacies generated by misconceptions of their interactions.

Her careful studies of the identification of psychological traits have encouraged the construction of instruments better suited for measuring these traits, to avoid the misleading generalities that have frequently characterized such measures.

Her respect for racial and cultural diversity underlies her concern for appropriate consideration of ethnic differences by test makers and test users.

Her excellence as a master teacher is reflected in the numerous editions of her classic texts, *Differential Psychology* and *Psychological Testing*. Her editorial skills are well documented in *Testing Problems in Perspective*, a collection of papers from the ETS Invitational Conference covering twenty-five years of progress in measurement. Her other books and a wealth of articles provide theorists and practitioners with technical and practical information on topics ranging from test construction and statistical methods to studies in creativity.

She has given freely of her talent for executive leadership in her presidencies of the American Psychological Association, the American Psychological Foundation, the Eastern Psychological Association, and other professional groups, and in her chairmanships of many committees and conferences.

Expressing the admiration and gratitude of the educational measurement community for her invaluable contributions to its body of knowledge and understanding, ETS has the honor to present the 1977 Award for Distinguished Service to Measurement to Anne Anastasi.

Previous Recipients of the ETS Measurement Award

1970 *E. F. Lindquist*
1971 *Lee J. Cronbach*
1972 *Robert L. Thorndike*
1973 *Oscar L. Buros*
1974 *J. P. Guilford*
1975 *Harold Gulliksen*
1976 *Ralph Winfred Tyler*

Session I

Presiding, WILLIAM W. TURNBULL

Equal Protection and External Screening: Davis, De Funis, and Bakke

BARBARA LERNER
Psychologist and Attorney

External screening is the process by which an institution decides which outside applicants to accept and which to reject. Such screening is a necessity whenever applicants outnumber openings. In America today, applicants outnumber openings for most jobs and for most graduate and professional school places. The inevitable result is that many are disappointed. The disappointed come in all colors and both sexes and they come from a wide variety of ethnic, religious, and socioeconomic backgrounds. No background creates indifference to rejection; none immunizes against the pain of disappointment and the sense of inferiority and/or injustice it can generate.

Screening may be inevitable; no particular method is. Institutions choose the screening methods they employ and their choices have an enormous impact on American society. Wise choices can decrease feelings of inferiority and injustice for all Americans, regardless of ancestry; unwise ones can increase them to dangerous levels. The thesis of this presentation is that American graduate and professional schools have made a series of unwise choices in recent years, as illustrated by the record in the *De Funis*[1] and *Bakke*[2] cases, and that the foundation for a better approach can be discerned in the *Davis* case.[3]

All three of the above-mentioned cases deal with the relationship between external screening and the Equal Protection Clause of the Fourteenth Amendment. All have reached the Supreme Court of the United States in this decade and all have excited an unusual amount of controversy and commentary. Instead of focusing directly on these cases, my plan is to try to illustrate the thesis set forth above by presenting the following: first, a brief history of the screening

procedures American universities have used over the past four decades; second, a challenge to the two principles that have come to dominate their approach to external screening in recent years; and third, a rough outline of what seems to me to be a better alternative.

I. External Screening for Admission to Graduate and Professional Schools: The Status Quo, Past and Present

The first external screening case to reach the Supreme Court, *De Funis v. Odegaard*, involved a charge by the plaintiff, a rejected white male applicant, that the screening procedures used to deny him admission to the University of Washington Law School violated his right to equal protection by discriminating against him because of his race. Law schools are hardly the only institutions whose screening methods are being challenged these days. The challenged institution in the Bakke case—just argued in the Supreme Court—is a medical school. In the *Davis* case, it was not an educational institution at all, but a police department, deciding which job training applicants to accept and which to reject. Of course, external screening methods vary within and between groups of institutions, but overall, the approaches of American universities have been similar enough to permit one group of institutions, law schools, to serve as a useful illustration of what most of the others have done in the past and are doing now.

Initially, American law schools needed no external screening methods and used none. As recently as in the 1940s, the number of applicants did not exceed the number of openings. Almost anyone with a bachelor's degree who wanted to enter law school could and did. The trick then was to stay in, and it was a hard trick. Flunk-out and drop-out rates of over 50 percent were common.[4] This method—"throw them all in and let them sink or swim"—was harsh, particularly on applicants who entered the pool wearing lead shoes but did not discover that painful fact until after a large and fruitless expenditure of money, time, and effort. In addition, it was an inefficient and expensive method for the institutions involved in terms of wasted resources, lowered morale and performance levels, and high training and turnover costs. The only advantage to this no-

screening system was that everyone who wanted a chance to try got it.

In the fifties, the situation began to change. Increasingly, the number of applicants came to exceed the number of openings, and some form of external screening became a necessity. Law schools responded in the most reasonable way possible. They adopted a test admirably designed to accomplish its purpose: to predict in advance which applicants had a reasonable chance of successfully completing the course of studies necessary to prepare them to function as competent attorneys and which did not. The test was the LSAT[5] and, as a result of its widespread use, flunk-out and drop-out rates declined dramatically.[6]

What the law schools were doing at this point, in effect, was screening on the basis of necessary ability. The ability in question, verbal-reasoning skill, is obviously not the only ability necessary for satisfactory performance as a law student, attorney, or judge, but it is, undeniably, a necessary one—a "key element," to use more technical jargon—and it is the only relevant ability that is now relatively easy to measure in an objective fashion. This type of screening involves the use of an absolute cut-off score to divide those who have the necessary ability from those who lack it. I call it absolute because it is a single, fixed point, a constant dividing line chosen to represent a necessary ability level for a given task. It is, of course, not an absolute in the sense of being either preordained or infallible.

Selecting a cut-off point is, in essence, a matter of choosing the odds one wants to play by. Thus, if research with the LSAT shows that applicants who score above point A have only a 50–50 chance of successfully completing the program at a particular law school whereas applicants who score above point B have a 70 percent chance and applicants above point C a 90 percent chance, one is free to set the cut-off at any of those three points. As a result, administrators who adopt this approach are not donning strait-jackets or endorsing a minimal standard. Considerable latitude for discretion remains.

The main limit is that perfection is not attainable. Any cut-off score will exclude some who might have made it in spite of the odds against them and will include others who will not make it in spite of the odds favoring them. In general, law schools chose to set their cut-off points at fairly high levels in order to maximize the number of

successes and minimize the number of failures. The results, overall, were positive ones, for the schools, the applicants, and society as a whole.

Moving from an Absolute to a Relative Standard

Society, however, kept on growing, in terms of size and aspiration levels. As both the population and the percentage of that population wishing to become lawyers increased,[7] law schools faced a new problem: what to do when there were 2,000 applicants for 500 seats and 1,000 of them scored above even the high cut-off points established. What the law schools did was to move from an absolute to a relative standard.

What that means is that instead of deciding how high a score was necessary for competent performance in law school and equalizing the chances of everyone who scored above that point, law schools chose to differentiate not only between qualified and unqualified applicants but also between the qualified ones, selecting only those with the highest scores and rejecting those with lower, but still quite adequate, scores. Thus, instead of striving to meet a fixed standard, qualified applicants to law schools are now forced to try to score above a constantly shifting point in order to beat out other qualified candidates. How high is high enough depends upon how high one's competitors manage to score in the particular year in which one happens to apply.

Later, an attempt will be made to analyze the consequences of this approach and to compare them with the consequences of other possible approaches. For present purposes, it will suffice to illustrate the inflationary impact of a relative standard over time. The LSAT has a maximum range of 800 points. In 1961, the median LSAT score of students at 81 percent of the nation's law schools was below 485.[8] In 1975, not one of 128 ABA-approved law schools had an entering class with a mean below 510. Seventy percent of them had means between 572 and 693.[9] What this means in comparative terms is that most American lawyers and judges practicing today would never have gotten into law school at all if they had had to compete against the inflated standards which now govern admission.

This single, inflationary, relative-standard approach was used by

virtually all American law schools throughout the sixties. One result was that large numbers of qualified candidates of all races were rejected and made to feel intellectually inadequate, even though, in fact, they were not. Another was that among the qualified candidates who were accepted as of 1969, only 3.8 percent were black.[10] This low percentage embarrassed American law schools and, in the seventies, they chose to increase it by shifting to a dual relative-standards system.[11] They did this by segregating black and white applicants and accepting only the highest scorers within each separate racial group. Because the median score within the white group was higher than that within the black group, this meant that many rejected white applicants had scores above those of many accepted black applicants.

Marco De Funis, a would-be law student, and Allen Bakke, a would-be medical student, were both in that position and they, like thousands of other applicants in the same position, felt unjustly treated. Their sense of injustice is widely shared: A recent Gallup Poll showed that 83 percent of the population felt that minorities should not be given special preferences.[12] More striking still is the fact that 64 percent of nonwhites agreed.[13] Nonetheless, most American universities have adopted a dual-standards approach which was superimposed upon their earlier choice of a relative rather than an absolute merit standard. The result is a system characterized by two principles: duality and relativism.

Before concluding this description of the status quo and going on to challenge it, there is a ritual gesture that I am supposed to make: In addition to noting that the LSAT and other admissions tests are not used alone but in conjunction with another objective numerical indicator of ability—the UGPA, or undergraduate grade-point average—I am supposed to say something like this: "Of course, universities should not rely exclusively on objective numerical indices. There are numerous unquantified subjective factors which also affect an applicant's performance in school and thereafter, and these, too, rightfully enter into admissions decisions." As you have doubtless gathered, I am not comfortable about making this gesture.

It is not that I wish to derogate the importance of subjective factors. Far from it. John Dean is a more recent law school graduate than either of the Watergate special prosecutors, Mr. Cox and Mr. Jaworski, so it is probable that his LSAT scores were higher, but few among us would regard him as the better lawyer. Integrity, courage,

compassion, wisdom, and judgment seem to me, as I believe they do to most Americans, to be at least as important as the purely intellectual abilities.

My problems with the ritual gesture are twofold. First, it seems to me to be somewhat disingenuous. Few admissions officers really look at subjective factors affecting those who score below some inflated quantitative qualification line, whether there be one line for everybody or one for whites and another for blacks. Second, I am troubled by the widespread tendency to confuse subjective factors with subjective methods of assessing them. In psychometric theory, there is no necessary relationship between the two: Subjective factors can and should be objectively identified, defined, and assessed—even quantified. Nothing we now know suggests that that is an inherently impossible task. It is only a very difficult one, requiring sustained and systematic effort over a long period of time, and the sad fact is that as yet, we have not really expended that sort of effort in this area. Instead, what we have done is to let various collections of officials decide the fate of large numbers of qualified young people on the basis of personal feelings, hunches, and biases in ways that are often startingly unsystematic, inconsistent, and arbitrary, as the record in the *Bakke* case seems to illustrate.[14]

Arbitrary standards run the risk of violating the due process clause as well as the Equal Protection Clause.[15] In addition, they are a confession of ignorance, showing that we do not yet know how to identify and measure relevant subjective factors. Worse still, they are a guarantee of continued ignorance. Unless we are explicit about our hunches and systematic in applying them and checking the results, we will never learn which ones are valid and which totally invalid. That, and not numbers per se, is what psychometric testing, and science itself, is all about. It is a method that permits us to learn from our mistakes so that we do not make the same ones over and over again.

II. Challenges to the Status Quo

The Case for Abandoning Dual Standards

Basically, the argument for dual standards is that they are necessary because very few minority students would be admitted without them

and that would be a bad thing for three reasons: 1) because it would deprive all students of the educational benefits of diversity; 2) would lessen our ability to provide adequate professional service to all groups; and 3) would be unfair to the disadvantaged. In the paragraphs which follow, I am going to try to convince you that all three of these fine-sounding arguments are spurious.

Let me start with diversity. As a first-generation American who grew up in a multi-ethnic urban neighborhood and felt enriched by the diversity around me, I have a high opinion of its value. I also have some ideas about how to achieve it without using any quotas or giving special preferences to any group, ideas I shall set forth in the last section of this paper. For now, the relevant point is that very few proponents of dual standards actually seem to know or care much about diversity, about the truly remarkable variety of heritages which make the American experiment unique.

Typically, proponents of dual standards refuse to recognize the fact that we are a nation of Poles, Italians, Irish, Greeks, Armenians, Lithuanians, Germans, Dutch, and English, to name only a few.[16] Instead, they insist upon lumping all of these diverse peoples together into one category which, happily, still seems artificial to most Americans who are not members either of the Ku Klux Klan or of the American Nazi Party: white people.

Diversity is, in fact, used as a euphemism or code word for a handful of currently favored minorities. I use that cumbersome phrase because I know of no other way to characterize the groups singled out by EEOC and included in most affirmative action programs.[17] In theory, there are four; in practice, three, or more accurately, two and a half. The theoretical four are these: blacks, American Indians, Spanish Americans, and Orientals. The practical two and a half are blacks, Indians, and some Spanish Americans but not others—mainly those from Mexico and Puerto Rico but not those from Spain, Cuba, or other Latin American countries. The latter Spanish Americans are excluded for the same reason that Orientals were excluded from the special admissions program challenged in the *De Funis* case:[18] Their LSATs and UGPAs were high enough to permit them to obtain adequate representation even in the face of the grossly inflated relative standards which govern law school admissions today.

Whether one looks at the theoretical four or the practical two and a half, it is inaccurate to characterize them as nonwhite peoples. Ac-

cording to U.S. Census statistics, 98 percent of all Spanish Americans are white.[19] Are they then distinguished by the fact that as groups, they generally score lower than members of other groups on tests? Not really, even if one ignores the problem posed by the high-scoring Japanese and other Orientals such as the Vietnamese. Earlier studies, most of them badly in need of updating,[20] suggest that Japanese, like Jews, tend to have higher group averages on many tests than White Anglo-Saxon Protestants but to the best of my knowledge, no affirmative action official has yet suggested that WASPs, with the possible exception of Appalachians, be included in any special admissions program. Similarly, such statistics as we have suggest that as groups, Italians, for example, may tend to score below WASPs, but they are not included either. A history of discrimination does not work any better: There is no evidence proving that prejudice against Puerto Ricans is any greater than that against Greeks, Turks, or Arabs, for example.

One could go on and on with this listing of nonprinciples, if time permitted. It does not, so I will simply stop and go on to the next argument for dual standards. I will also continue to use my own euphemism for currently favored minorities—*blacks*—partly in the interest of brevity and partly because, like law schools, blacks provide a useful illustration of a more general phenomenon.

Argument 2 has to do with the need to provide more adequate professional services. This, too, is a code phrase requiring translation. What it really means is "we need more black lawyers" and, on the surface, it seems a powerful argument. Thus, given the following statistics and asked: "Should there be more black lawyers?" most men and women of good will would probably answer "yes." The figures are these: Blacks constituted about 11 percent of the total population and 12.7 percent of those in the 21-to-25-year-old age range usual for law school entrants, but only 5.3 percent of all law students, in 1976.[21] Prior to the widespread institution of dual standards in the 1970s, they constituted an even smaller proportion of all law students: 1.3 percent in 1964[22] and 3.8 percent in 1969.[23]

What these figures actually prove is that it is misleading to think about any problem in a vacuum, divorced from its context. In context, the question is not "Should there be more black lawyers?" but "Compared to what?" Please note that the question here has to do with *what*, not *who*. The *who* has already been decided: Only 5.3 percent of all college graduates are black,[24] and it is this 5 percent

that the law schools and medical schools are fighting over. If they should lose the fight, poverty and unemployment will not win it; other graduate schools and/or professions, such as physics, chemistry, history, education, engineering, journalism, social work, and business administration,[25] will.

Whichever discipline wins, none of the 5 percent are likely to end up as impoverished, unskilled laborers. Neither are they likely to get rich by going to law school. CLEO, the Council on Legal Educational Opportunity, a group devoted to recruiting black college graduates for the law, reports that their average black law school graduate earned between $10,000 and $12,000 at his first post-law school job.[26] CLEO officials shrink from admitting the fact that white law school graduates earn considerably more than that, but go on anyway to explain the discrepancy. They suggest that black attorneys prefer to work in criminal and civil rights law serving the poor rather than in more lucrative areas like corporate finance and taxation.[27] Perhaps, although that is pure speculation without a shred of supporting evidence.

Robert L. Tucker, the black attorney for the Reverend Jesse Jackson's Operation PUSH, offers a different explanation. He concludes that black lawyers in general are not making big money because there still aren't enough black-owned big businesses to support black law firms.[28] In the regrettable absence of conclusive evidence for either position, Mr. Tucker's speculations seem at least as worthy of serious consideration as do those of CLEO officials. In either case, it is far from clear that black college graduates—or the black community as a whole—will derive greater benefits from careers in law than from careers in business, the sciences, and the arts, and that—when one cuts through all the rhetoric and the De Tocqueville quotes—is what the fighting on this issue is really about. So much for argument 2.

Argument 3, that dual standards are necessary if we are to be fair to the disadvantaged, is even weaker. Disadvantaged students are those from homes characterized by poverty and/or ignorance, and it is as insulting as it is inaccurate to equate black with poor and ignorant. The average black American today is neither. In fact, although a greater percentage of blacks than of whites fall into the impoverished and undereducated category, two thirds of the unfortunates in that melancholy position are white.[29] Such evidence as we have suggests that what is true for the population in general is

also true for law school applicants.

In 1976, those applicants were asked this question: "Would you describe yourself as coming from a low income family background, such as from a family with a yearly income under $6,500 during your pre-college years?"[30] Three-quarters of the "yes" answers came from whites.[31] If American law schools were to give preferential treatment to all students who answered "yes," regardless of ancestry, no court would be likely to find their action unconstitutional. That, however, is not what the law schools have done. Instead, they have given preferential treatment to blacks from wealthy and well-educated families while denying it to whites from poor and poorly educated families.

Disadvantaged, it seems, is only another euphemism for currently favored minorities. The sole justification advanced for this sort of help-the-rich-at-the-expense-of-the-poor policy is that blacks suffered from segregation, whites did not. This was certainly true, and utterly deplorable, prior to 1954 when *Brown v. Board of Education*[32] was decided. However, this happens to be 1977, and from this, it would seem to follow that the average 21-year-old black applicant was born in 1956 and entered first grade in 1962, eight years after *Brown* was decided.

By that time, state-enforced segregation was largely a thing of the past. What remained then and continues today is a high degree of voluntary separation between all groups with distinctive heritages. Then, as now, many Italians—to choose one example almost at random—preferred to live in dominantly Italian neighborhoods, not because the state made them do it and not because they hated other Americans with different ethnic, religious, or racial backgrounds, but because they valued their distinctive cultural traditions and did not want to be melted down in the great Zangwillian pot or otherwise divested of their identity.[33] Neither were they rejecting the American dream. Rather, they, like countless other groups of Americans, felt that they could make a positive contribution as Italian-Americans without sacrificing either the hyphen or the word that preceded it.

State-enforced segregation was an unmitigated evil because it wrongfully deprived black Americans of the choice between assimilating and maintaining a separate group identity. Government-enforced integration runs the risk of again depriving black people of their right to choose. My own view is that the government has no

more right to force integration on blacks than it had to force segregation on them. Recent decisions suggest that the current Supreme Court, while remaining absolutely unyielding in its opposition to state-enforced segregation and discrimination, does not view the Equal Protection Clause as a mandate to homogenize the country.[34]

The NAACP may be distressed because it has always been strongly in favor of integration and assimilation, but many blacks do not agree.[35] The point here is not that the NAACP is wrong and more ethnically oriented blacks are right. Rather, the point is that this is a choice that each black man and woman should have a right to make for him- or herself, free of extraneous legal pressures. The blind insistence on condemning as inferior the education of every black child who did not go to a white-dominated school is one such extraneous pressure.

In fact, such evidence as we have suggests that most blacks are more concerned about quality education than they are about integration and that they refuse to equate the two.[36] Since we have no real evidence to indicate that they are wrong[37] and no right to make the choice for them, perhaps it is time we stopped telling them that their education was inferior whenever most of their classmates were black. If we do stop, then there is no basis for preferential treatment for blacks *qua* blacks. There is a basis for preferential treatment for poor blacks as there is for poor whites: fairness to the disadvantaged in the literal, not the euphemistic, sense.

The arguments for dual standards may be weak, as I have tried to demonstrate; the arguments against them are not. First and foremost, dual standards compromise the constitutional principle that discrimination on the basis of race violates the Equal Protection Clause. Counter arguments about benign purposes or effects miss the point. As the Supreme Court explained in *Brown*, "In the first cases in this Court construing the Fourteenth Amendment, decided shortly after its adoption [in 1868], the Court interpreted it as proscribing all state-imposed discrimination against the Negro race."[38] The basic principle—that a person's constitutional rights cannot depend on his color—was not compromised in any way until 1896 when the Court first permitted state-imposed discrimination on the basis of race, rationalizing its retreat with the "separate but equal" doctrine.[39]

The damage done by that single qualification of principle was, and still is, incalculable. It took more than half a century to recover the

legal high ground that had been lost. Slowly, painfully, beginning with the 1938 case of *Missouri ex rel Gaines v. Canada*,[40] the Court retraced its steps until finally, in the *Brown* case, it removed the last taint of qualification, holding that "[I]n the field of public education, the doctrine of 'separate but equal' has no place."[41] To then turn around, only a little more than two decades later, and allow discrimination on the basis of race, for any purpose, is to compromise that principle once again, to add needless complications, qualifications, "balancing tests," ambiguities, and loopholes which will haunt us ever after.

Even the strongest proponents of dual standards concede that at best, they can only be a temporary expedient.[42] The weakening of a key constitutional principle, however, would be a permanent loss. Much more could and should be said about this cardinal legal issue. Fortunately, that has been done already, and so masterfully that the briefest and best advice I can give is to urge everyone to read the amicus briefs by Philip Kurland and his colleagues in the *De Funis* and *Bakke* cases.[43]

Two other arguments against dual standards are worthy of serious consideration. First, their use heightens the community's sense of injustice and, as shown by the Gallup Poll results cited earlier,[44] the community in question includes substantial majorities of all citizens, black and white. In light of that fact, recipients of special preferences are doubly stigmatized, not only as persons of inferior ability but also as persons who are taking unfair advantages. Worse still, because the basis for the special preference is race, the stigma inevitably attaches not only to blacks who received it but also to the many blacks who did not and achieved their goals without benefit of any special preferences whatsoever.

The resultant risk of simultaneously reinforcing black feelings of inferiority and white feelings of superiority is so great that many thoughtful black people are appalled and outraged. Kenneth Clark, the black psychologist whose pioneering research was cited in the famous footnote 11 to the *Brown* decision, put it this way: "For blacks to be held to lower standards or in some cases no standards at all is a most contemptible form of racism."[45]

The Law School Admission Council has a rather remarkable answer for Professor Clark and others like him. In their amicus brief in the *Bakke* case, they argue that there should be no stigma because the black applicants admitted under the separate lower standard are

no less qualified than the black and white applicants admitted under the higher regular standard. The analysis on which their argument is based seems to me, as I believe it would to Professor Clark or to any other competent psychologist, virtually incontrovertible. The problem is that the conclusion they draw from that analysis is a total non sequitur, devoid of logic, fairness, and rationality.

Here, in slightly abbreviated form, is their two-part analysis. First, they tell us that UGPAs and scores on the LSATs allow us to predict potential law school failures with a very high degree of accuracy "in the lower ranges of the applicant pool" and that "the use of these predictors for this specific purpose is therefore vital and abandonment would be foolhardy."[46] Then they go on to the second part of their analysis. Let me quote the entire paragraph verbatim:

> Well above the range of probable failure, however, lies a much larger volume of applicants than the schools' total capacity. All are fully qualified to perform well on law school grades, and many are nearly indistinguishable on these measures. In this range, where most of the admissions work must be done, predictions of relative law school ranks are less accurate. But at the same time, they are less significant. Whether an applicant is predicted for the 40th or 50th percentile of the class is a matter of no real consequence.[47]

Bravo! A flawless analysis proving that it makes no sense to reject competent blacks on the basis of a relative standard so ludicrously inflated that it serves only to predict, weakly, "a matter of no real consequence," a standard that, in its upper reaches, is well nigh meaningless. This is hardly a justification for the adoption of dual standards. Rather, it is a compelling argument for the abandonment of inflated relative standards for all applicants, white and black. It is to that topic that I turn next.

The Case for Abandoning Inflated Relative Standards

There are three main arguments against the use of inflated relative standards. The first is that they bear no meaningful relationship to intellectual merit and may be inversely related to important, nonintellectual traits which are also necessary for competent performance in the law as in other fields. The point about the lack of relationship between intellectual merit and test scores above the necessary

ability level has already been made in the paragraph from the LSAC brief quoted above. In less technical language, the point is that test score differences above a certain level become, in essence, differences between Tweedledum and Tweedledee. In this realm, the maximum is not necessarily the optimum. More simply still, "more" is not the same as "better," and the simple-minded insistence on regarding it as such is potentially quite harmful and counterproductive.

This is so, as I tried to explain in an earlier paper,[48] because while a certain level of verbal facility and reasoning ability may be necessary for competent performance as a lawyer or anything else, above that level the differences between superior and inferior workers are a function of other characteristics. Focusing students' energies on the meanly competitive task of scrounging about for an extra half-point advantage on an exam can be demeaning and demoralizing. It is conducive to the development of a fierce but intellectually contentless competition at the expense of intellectual depth, originality, or character. Ex hypothesi, it may well serve to eliminate from the upper echelons of the professions many people of superior ability who pursue intellectual enlightenment rather than points and, as a result, earn scores or grades which, on the average, could be predicted to be somewhat lower than those earned by the single-mindedly opportunistic.

A second argument against the use of inflated relative standards is that screening on this basis produces unjustified feelings of inferiority in hundreds of thousands of people whose abilities are fully adequate for the positions they aspire to but who are led to believe that they have been rejected because their abilities are not good enough. Telling them that the standards by which they have been judged and found wanting have become so arbitrary as to verge on meaninglessness would be more candid but is unlikely to mollify them. Neither is it likely to help them understand and deal with the late twentieth century world they live in and the special problems it has produced, problems which revolve around the fact that there are too many people and too few places for them.

The third argument against the use of inflated relative standards is that such an approach provides the basis for what has become one of our most dangerous and least rewarding national pastimes: making endless, artificial comparisons of the abilities of whites versus blacks. Please note that I am not—repeat, not—challenging the

need to make real comparisons on any meaningful basis. If the children of the poor are not learning the basic skills necessary to give them a real vocational chance and some real vocational choices, we need to know that fact so that we can work to change it. Similarly, if the black poor are learning even less than the white poor, we need to know that too—again, in order to change it. As my colleagues in psychology have repeatedly pointed out, it is as senseless to attack valid tests which reveal genuine deficits as it is to try to break "a thermometer because it registers a temperature of 101°."[49] Killing the messenger who brings bad news is just not an effective problem-solving technique.[50]

What I am challenging are comparisons that focus on trivial differences which are just that: differences, not deficits requiring remedies. Asking whether a greater percentage of whites than of blacks obtain LSAT scores over 700 is like asking whether more whites or more blacks are over seven feet tall. I do not know the answer to that question but if I did, I would put it in a trivia collection because that is where I think such information belongs.

Instead, such trivia has become a basis for national handwringing, weighty policy decisions, and large-scale guilt and inferiority trips and is now put forth as a reason for eviscerating a fundamental constitutional guarantee. All this, despite the fact that an applicant, black or white, can have an LSAT score literally hundreds of points below that and still not have anything that could fairly be regarded as an intellectual deficit for the study of law or anything else. The purported need for double standards bears little, if any, relationship to black deficits; it is largely a function of inflated relative standards. A return to reasonable standards would therefore obviate the need for double standards; it is, in essence, as simple as that.

These, then, are the arguments for the abandonment of inflated relative standards. To the best of my knowledge, there are only three arguments for the retention of such standards: 1) they insure constant progress; 2) they maximize competition; and 3) anyway, there is no alternative. Regarding the first argument, at this point, it should suffice to say that movement per se, up or down, is neither an infallible sign of progress nor an omen of retrogression. The real issue here is not progress but prestige: Schools with the highest relative standards may not really have better students and may not really produce better lawyers but they do have the most prestige, and many individuals have a heavy investment in that prestige

system. Still, it hardly seems worth the social and constitutional cost, especially since there are other, better ways to maintain a prestige system and to keep it related to real rather than artificial merit.

Quality of faculty is one time-honored alternative and a good one. Thus, if the Yale Law School had abandoned relative standards yesterday and equalized the admission chances of every applicant with an LSAT score of 500 or more, for example, it would still have been regarded by many as an especially desirable place to study, if for no other reason than because the late Alexander Bickel was there.[51] Many admirers of the work of Philip Kurland, myself included, feel the same way about the University of Chicago.[52] Paul Freund of Harvard[53] and Herbert Wechsler of Columbia[54] are similarly regarded by many, and for similarly good reasons.

The second argument for relative standards is that they maximize competition and, while it is certainly clear that they do that, it is much less clear whether that is really an absolute good, in and of itself. John Dean would probably get very high marks on a scale of sheer competitiveness, but that seems to be at least as much a part of the problem as it is of the solution.

Perhaps because the foregoing arguments are so weak, most amicus briefs in the *Bakke* case which defended the current relative dual-standards system did so by relying almost entirely on the third and last argument for it: There's no alternative.[55]

III. An Alternative: Building on Mayor Washington's Foundation

Since I have claimed, throughout this presentation, that there is a better alternative, I had best use what little time remains to at least sketch in the main contours of that alternative. What I am recommending is that American universities start afresh by abandoning relative standards and dual standards and returning to a single, absolute, necessary-ability-level standard for everyone. That was the approach adopted by Walter Washington, the black mayor of the District of Columbia, and his police chief, Jerry V. Wilson, to screen applicants for jobs as police officers.

They, too, used a test of verbal and reasoning ability, Test 21, but instead of encouraging endless, senseless, spiralling competition,

they used a single, absolute cut-off score chosen to reflect the level of ability necessary for competent performance as a police officer. They then accepted all applicants who scored above that line and rejected all applicants who scored below it. By combining this reasonable screening method with an affirmative-action recruitment program, they were able to hire large numbers of competent police officers. No double standards and no racial quotas were used and none were needed: In the initial hiring period, 44 percent of all new officers hired were black.[56] Later, that figure rose to 57 percent.[57]

This was the screening system challenged in the *Washington v. Davis* case and upheld by the Supreme Court as consistent with the Fourteenth Amendment.[58] It is the system I believe American universities should also adopt and build upon. Officials of many such institutions will immediately protest: "Impossible! We can't accept all qualified applicants; there just aren't enough openings." There is a simple answer to that: Use current tests like the LSAT to separate the qualified from the unqualified and then equalize the chances of every qualified applicant by selecting from among them on a random basis.

Randomization has a great many advantages over the present university selection system. I will very quickly list the more important ones and then conclude this presentation with some suggestions about how randomization can be coupled with other methods in ways that enable us to build on the Washington foundation in order to reduce our ignorance and increase the efficiency and fairness of our screening systems.

The main advantages of randomization are these: First, it would maximize diversity, not just between racial groups but within them, and without using quotas and the insoluble problems inherent in their use in a society as heterogeneous as ours. Second, it would reduce the community's sense of injustice: No group could be accused of having an unfair advantage over any other group. Third, it would curtail the spread of unwarranted feelings of inferiority: The luck of the draw rather than lack of ability would become an acknowledged basis for many rejections. Fourth, it would reduce destructive and pointless competition and hostility between individuals and groups. Fifth, and perhaps most important, it would focus everyone's attention on the real problem: the fact that we have too many people and too few places.

Why that is so and what can and should be done about it is a topic

worthy of extended discussion in its own right, but one that is beyond the scope of this paper. Here, the focus is on dealing with the problem as it currently exists by improving external screening methods. Wiser use of the many good tests of intellectual abilities we already possess will solve some of our immediate problems. Further progress, however, depends upon our learning how to identify and measure other relevant abilities so that instead of screening for only one necessary ability or quality, we can also screen for other, equally important ones. To do that, we must be able to test new measures of additional qualities, discarding those that prove invalid and retaining those that prove valid.

American universities are already experimenting with a variety of additional screening ideas, using subjective factors to select from among an artificially narrowed group of extremely high scorers on current measures. Such experimentation is inevitable; it cannot and should not be eliminated but it could and should be done in a more systematic, scientific fashion so that it will teach us more than it currently does.

If, instead of using the same unvalidated, second-stage screening methods on all qualified applicants, universities were to select at least half of them at random and the other half on the basis of whatever new factors or measures seem promising to them, we would be in a better position to learn from our mistakes and in less danger of repeating them. This happier outcome would be especially likely if universities were to then commission careful, long-term follow-up studies of how the members of each selection group perform, not only in school but afterwards as well. Because randomization would largely obviate the restriction-of-range problem which has dogged researchers in this area, the chances of successfully identifying valid new predictors would be markedly improved.

If such procedures are followed, we should, eventually, learn how to screen in ways that genuinely foster and reward real merit for all of our citizens. The goal, in sum, is to get all of us off the trivia treadmill, redirecting our energies into the struggle to achieve a true meritocracy for the benefit of all of our citizens.

Footnotes

1. De Funis v. Odegaard, 416 U.S. 312 (1974).
2. Regents of the University of California v. Bakke, Supreme Court of the United States No. 76-811, October Term, 1976.
3. Washington v. Davis, 426 U.S. 229 (1976).
4. L. NICHOLSON, THE LAW SCHOOLS OF THE UNITED STATES 26 (1958).
5. *See e.g.*, Schrader & Olsen, *The Law School Admission Test as a Predictor of Law School Grades*, Report #LSAC-50-1, Law School Admission Council. REPORTS OF LSAC SPONSORED RESEARCH: VOL. I, 1949-69. French, *Validation of the Practical Judgment and Directed Memory Experimental Sections of One Form of the Law School Admission Test*, Report #LSAC-52-1, *id;* Johnson & Olsen, *Comparative Three-year and One-year Validities of the LSAT at Two Law Schools*, Report #LSAC-52-2, *id. See also* Brief Amicus Curiae for the Association of American Law Schools in the Bakke case, 10, 48: "Predictions of Law School Grades are Approximations, Useful Originally to Exclude Probable Failures."
6. The overall law student attrition rate was 38.4 percent in 1950. It declined to 29.8 percent in 1960 and to 12.7 percent in 1970. The inflated standards of the 1970s produced no comparable effect: Attrition rates in 1975 were 10 percent. *See* ABA LAW SCHOOLS AND BAR ADMISSION REQUIREMENTS (1950)-(1976). *See also* Brief Amicus Curiae for the Association of American Law Schools in the Bakke case at 58.
7. An especially dramatic growth spurt took place in the late 1960s and early 1970s: 50,793 persons took the LSAT in 1967-68; 121,871 did so in 1971-72, an increase of 140 percent in four years. During that same four-year period, the number of first-year law students also increased, but by "only" 47 percent, from 25,746 in 1967-68 to 37,724 in 1971-72. *See* Linn, *Test Bias and the Prediction of Grades in Law School*, Report #LSAC-75-1, LAW SCHOOL ADMISSION RESEARCH, 1976. The situation has remained essentially unchanged since then: There were 133,000 LSAT administrations in 1975, but only 39,038 first-year seats in ABA-approved law schools. *See* Brief Amicus Curiae for the Association of American Law Schools in the Bakke case at 11.
8. S. WARKOV, LAWYERS IN THE MAKING, 1965.
9. Evans, *Applications and Admissions to ABA Accredited Law Schools: An Analysis of National Data for the Class Entering in the Fall of 1976*, Report #LSAC-77-1, LAW SCHOOL ADMISSION RESEARCH,

1977 at 5, hereafter cited as *Evans Report*. These comparisons, it should be noted, may understate the magnitude of score inflation because the Warkov study reports medians, while Evans reports only means. Some notion of the difference this can make can be gleaned by perusing the 1972–73 PRELAW HANDBOOK prepared and published by the Association of American Law Schools and the Law School Admission Council. Only 14 of the 143 law schools described therein listed the median LSAT score of their 1971 entering class, but two of those, Chicago and Yale, had overall medians above 693, the maximum mean score for schools in the top 10 percent in the 1975 *Evans Report*. Chicago's 1971 median was 697, Yale's was 723, and the Yale entry indicates that the listed median was depressed by the inclusion of LSAT scores of minority students admitted under their special admission program. The median LSAT score of those minority students was 683 in 1971. Because LSAT scores have tended to rise each year for the past several years, the odds are great that current medians are higher still, and that the overall mean-median discrepancy is even larger.

10. *Evans Report* at 7. *See also* Report of Minority Groups Project in American Association of Law Schools Proceedings 172 (1965). In the latter source, blacks were estimated to constitute only about 1.3 percent of all law students in 1964 (roughly 700 out of 54,265).

11. *See, e.g., Symposium, Disadvantaged Students and Legal Education—Programs for Affirmative Action*, 1970 TOL. L. REV. 227. See also Brief Amicus Curiae for the Association of American Law Schools in the Bakke case at 24. "Thus, by the mid 1970s, in virtually all schools, in one way or another, a preference in the application of admission standards was in fact afforded to applicants from minority groups."

12. *See* The New York Times, Sunday, May 1, 1977, p. A. 33, col. 1.

13. *Id.*

14. *See* Amicus Curiae Brief of the National Conference of Black Lawyers in the Bakke case for a detailed examination of some of the grosser inadequacies in the data provided by admissions officers of the University of California Medical School at Davis on criteria used to admit students through their regular and special admissions programs. One need not accept the conclusion of the brief's authors—that the university's position is not really adverse to Bakke's—to share their concern over what appears to be a lack of clarity and/or candor about how criteria were defined and applied. Such concern seems well warranted in light of available evidence on the invalidity of subjective methods of evaluation. *See* Linn & Winograd, *New York University Admissions Interview Study*, Report #LSAC-69-2, Law School Admission

Council. REPORTS OF LSAC SPONSORED RESEARCH: VOL. I. 1949–69.

15. *See, e.g.*, Pinsker v. Pacific Coast Society of Orthodontists. 12 Cal. 3d 541 (1974) for a recent statement that due process requires admission standards to be reasonably certain and consistently applied.

16. The same point has been made clearly and forcefully in several other recent articles. *See, e.g.*, Lavinsky, *De Funis v. Odegaard: The Non-Decision with a Message*, 1975 COLUM. L. REV. 520.

17. *See* EEOC Standard Race/Ethnic Categories. 41 Fed. Reg. 17601-02, (April 27, 1976).

18. *See* De Funis v. Odegaard. 416 U.S. at 338 where Mr. Justice Douglas, dissenting, noted that the special admissions program at the University of Washington Law School "included Filipinos, but excluded Chinese and Japanese." Japanese-Americans were also eliminated from the special admissions program at the University of California Law School when the faculty found that members of that group were being admitted in substantial numbers through the regular admissions program. *See Report on Special Admissions at Boalt Hall After Bakke*. 28 J. LEGAL ED. 363 (1977).

19. COUNTING THE FORGOTTEN. U.S. Commission on Civil Rights 43 (1974).

20. For an account of the largely unsuccessful efforts of non-Hispanic white ethnic Americans to get appropriate governmental agencies to compile current statistics on how members of their groups are faring in contemporary society, *see* Brief Amicus Curiae of the Polish American Congress *et al.* in the Bakke case. *See also* Handlin. *The Goals of Integration*, in PARSONS & CLARK. THE NEGRO AMERICAN (1967): "[T]he limitations of the census categories which recognize only whites and nonwhites obscure the genuine differences in occupation and income among the former and make comparisons invidious" (at 699).

Although the *Evans Report* also tries to lump all non-Hispanic whites together, some tip-of-the-iceberg indications of heterogeneity within that wastebasket category can be gleaned by examining the data provided in Appendix E on LSAT scores of three groups which are probably all or mainly white: those who identified themselves as white (48,249 applicants), those who refused to identify themselves (18,745 applicants), and those who identified themselves as "other" (2,152 applicants). Calculations based on these data show that 35 percent of the "whites" achieved LSAT scores ≥ 600, but 40 percent of the

unidentifieds and only 24 percent of the "others" did likewise. With regard to the reported discrepancies between black and "white" LSAT scores, it is also important to note that a larger percentage of black than of white college graduates apply to law schools. The result is that a less selected group of black applicants is competing with a more selected group of white applicants with a consequent exaggeration of the magnitude of intergroup differences.

More generally, historical data on the differential school achievement of various white ethnic groups are summarized in Greer, *Immigrants and School Performance*, in THE GREAT SCHOOL LEGEND (1972); and in RAVITCH, THE GREAT SCHOOL WARS (1974). For more recent data on the persistence of ethnic group differences, *see* Gross, *Learning Readiness in Two Jewish Groups*, Center for Urban Education (1967); Lesser, Fifer & Clarke, *Mental Abilities of Children from Different Social Classes and Cultural Groups*, Monographs of the Society for Research in Child Development, Vol. 30, No. 4 (1965); Majoribanks, *Ethnic and Environmental Influences on Mental Abilities*, 78 AM. J. SOCIOL. 323 (1972); Schwartz, *The Culturally Advantaged: A Study of Japanese-American Pupils*, in EPPS, ed., RACE RELATIONS (1973).

21. ABA Law Schools and Bar Admission Requirements: A Review of Legal Education in the United States, 42, 45 (1976).

22. Report of Minority Group Project in AALS PROCEEDINGS 172 (1965).

23. *Evans Report* at 7.

24. Atelsek & Gomberg, *Bachelors Degrees Awarded to Minority Students, 1973–74*, American Council on Education, Higher Education Panel Report No. 24 (1977).

25. *Minority Group Participation in Graduate Education*, National Board on Graduate Education (1976).

26. *See* Report on Survey of 1971–72 CLEO Graduates in Amicus Curiae Brief of the Council on Legal Education Opportunity in the Bakke case, 41–42.

27. *Id.* at 45.

28. *See* Chicago Daily News, Saturday–Sunday, August 6–7, 1977, p. 4.

29. Bureau of the Census, Current Population Reports, Series P-60, No. 103, Money Income and Poverty Status of Families and Persons in the United States: 1975 and 1974 Revisions (Advance Report, 1976).

30. *Evans Report* at 59.

31. *Id.*

32. 347 U.S. 483 (1954).

33. *See, e.g.*, H. GANS, THE URBAN VILLAGERS (1962); N. GLAZER & D. MOYNIHAN, BEYOND THE MELTING POT 181–216 (1963); PARK & MILLER, OLD WORLD TRAITS TRANSPLANTED 146–151 (1921). *See generally* KANTROWITZ, ETHNIC AND RACIAL SEGREGATION IN THE NEW YORK METROPOLIS (1973); Rosenthal, *Acculturation Without Assimilation*, 66 AM. J. SOCIOL. 275 (1960).

34. *See e.g.*, Milliken v. Bradley, 418 U.S. 717 (1974); Village of Arlington Heights v. Metropolitan Housing Development Corp., 97 S. Ct. 555 (1977).

35. *See, e.g.*, BRODERICK & MEIER, EDS., NEGRO PROTEST THOUGHT IN THE TWENTIETH CENTURY (1965); Handlin, *The Goals of Integration*, in PARSONS & CLARK, eds., THE NEGRO AMERICAN 659 (1967); ISAACS, THE NEW WORLD OF NEGRO AMERICANS (1963); MARX, PROTEST AND PREJUDICE (1967); WARREN, WHO SPEAKS FOR THE NEGRO? (1965). *See also* BRADBURN *et al.*, SIDE BY SIDE 133–34 (1971); Hauser, *Demographic Factors in the Integration of the Negro*, in PARSONS & CLARK, *op. cit.* at 96; WATTS, FREEMAN HUGHES, MORRIS & PETTIGREW, THE MIDDLE INCOME NEGRO FACES URBAN RENEWAL (1964).

36. *See* KELLAM, BRANCH, AGRAWAL & ENSMINGER, MENTAL HEALTH AND GOING TO SCHOOL 21 (1975); MARX, PROTEST AND PREJUDICE (1967); WATTS *et al.*, *op. cit.* at n. 35 *supra*.

37. *See, e.g.*, Coleman, *Toward Open Schools*, 9 PUBLIC INTEREST 20 (Fall 1967); Coleman, *Equality of Educational Opportunity: Reply to Bowles and Levin*, III J. HUMAN RESOURCES 22 (1968); Dyer, *School Factors and Equal Educational Opportunity*, 38 HARV. ED. REV. 53 (1963); MOSTELLER & MOYNIHAN, eds., ON EQUALITY OF EDUCATIONAL OPPORTUNITY (1972); ST. JOHN, SCHOOL DESEGREGATION: OUTCOMES FOR CHILDREN (1975).

38. 347 U.S. at 490.

39. Plessy v. Ferguson, 163 U.S. 537 (1896).

40. 305 U.S. 337 (1938). *See also* Sipuel v. Board of Regents, 332 U.S. 631 (1948); Sweatt v. Painter, 339 U.S. 629 (1950); and McLaurin v. Oklahoma State Regents, 339 U.S. 637 (1950).

41. 347 U.S. at 495.

42. *See, e.g.*, the following amicus curiae briefs in the Bakke case: Law School Admission Council Brief at 28; Association of American Law Schools Brief at 26, 27.

43. Both briefs were submitted on behalf of the Anti-Defamation League of B'nai B'rith. Collaborators on the De Funis Brief included the late Alexander M. Bickel as well as Larry Lavinsky and Arnold Forster. On the Bakke Brief, Messrs. Kurland, Lavinsky, and Forster were joined by Daniel Polsby and the following representatives: Leonard Greenwald for the Council of Supervisors and Administrators of the City of New York, AFSA, AFL-CIO, David Ashe for the Jewish Labor Committee; Dennis Rapps for the National Jewish Commission on Law and Public Affairs; and Anthony Fornelli on behalf of UNICO, the nation's largest Italian-American community service and public affairs organization.

44. *See* note 12 *supra*.

45. *See* The Chicago Tribune, June 29, 1971. *See also* Roy Wilkins, *The Case Against Quotas*, ADL BULL., March 1973 at 4; SOWELL, BLACK EDUCATION, MYTHS AND TRAGEDIES (1972); Sowell, *Affirmative Action Reconsidered*, 42 PUBLIC INTEREST 47, 63 (Winter, 1976).

46. Brief, p. 49.

47. *Id.*

48. Lerner, *Washington v. Davis, Quantity, Quality and Equality in Employment Testing*, 1976 SUPREME COURT REVIEW 263, 287.

49. ANASTASI, PSYCHOLOGICAL TESTING 60 (4th ed. 1976).

50. CRONBACH, ESSENTIALS OF PSYCHOLOGICAL TESTING 306 (3rd ed. 1970).

51. *See, e.g.*, BICKEL, THE MORALITY OF CONSENT (1975); THE SUPREME COURT AND THE IDEA OF PROGRESS (1970); POLITICS AND THE WARREN COURT (1965); *The Passive Virtues*, 75 HARV. L. REV. 40 (1961).

52. *See, e.g.*, KURLAND, WATERGATE AND THE CONSTITUTION (in press); THE PRIVATE I: SOME REFLECTIONS ON PRIVACY AND THE CONSITUTION (1976). *The Appointment and Disappointment of Supreme Court Justices*, 1972 LAW AND THE SOCIAL ORDER 183; MR. JUSTICE FRANKFURTER AND THE CONSTITUTION (1971); POLITICS, THE CONSTITUTION AND THE WARREN COURT (1970).

53. *See e.g.*, FREUND, THE SUPREME COURT OF THE UNITED STATES, ITS BUSINESS, PURPOSES AND PERFORMANCE (1961); *Storm Over the American Supreme Court*, 21 MODERN L. REV. 345 (1958).

54. *See, e.g.*, Wechsler, *Toward Neutral Principles of Constitutional Law*, 73 HARV. L. REV. 1 (1959); HART & WECHSLER, THE FEDERAL COURTS AND THE FEDERAL SYSTEM (1953).

55. *See e.g.*, Brief Amicus Curiae for the Association of American Law Schools in the Bakke case at 5: "The purpose of this Brief is to demonstrate a single proposition; the practice of providing a degree of preference for blacks and other minorities in law school admissions is a necessary, and indeed *the only* honest method, to achieve certain very important social objectives. Stated more bluntly, a holding that the Constitution requires that the schools abjure any consideration of race as a factor in making admissions decisions must, unless covertly circumvented, result in substantially all-white schools."

56. Davis v. Washington, 348 F. Supp. at 16 (D.C. Cir. 1972).

57. Davis v. Washington, 512 F. 2d at 961, n.32 (D.C. Cir. 1975).

58. Washington v. Davis, 426 U.S. 229 (1976).

Discussion

ERNEST M. BERNAL JR.
Associate Professor of Bicultural Bilingual Studies
The University of Texas at San Antonio

I find myself in a most peculiar position: While charmed by our speaker's pulsar presentation, I am compelled by my difference in perspective to don my black helmet, visor, and body armor and engage in some Star Wars, even at the risk of making her appear to be Princess Leia. Admittedly, Lerner wishes to expand her paper and is looking for a critique. I urge her to adjust her deflector shields, check the atomic batteries in her light sabre, and insert a fresh refill in her ballpoint pen.

To begin with, while I don't wish to dispute Dr. Barbara Lerner's personal claim to hold in high regard the ethnic diversity which she herself has experienced, I must say that her presentation demonstrates an almost singular insensitivity to cultural pluralism. Her credibility is shaken when she enumerates no less than nine Anglo minorities but insists on recognizing only four other minority groups under the rubric of their being "favored." Instead of seeing the cultural differences which exist within these four broad classifications, however, she promptly reduces them to two-and-a-half, and thereafter refers to these collectively as "blacks," apparently out of her concern for efficiency in communication. I would hope that our speaker, as a member of one of the minorities which constitute the privileged and educationally advantaged majority, would exhibit both *noblesse oblige* and ethnographic accuracy. Certainly no person conscious and proud of her or his minority culture appreciates being reduced to the status of shorthand.

It might be instructive for her to stop regarding us all as mere ethnics and to start thinking of us as ethnics who are either dominant or non-dominant, to consider whether a group's historical, cultural, or linguistic background either gives it access to the formal organizations or institutions of our society or becomes an impediment to its members' participation. In this way, she would not be tempted to

confuse numbers with proportions when discussing poverty and would gain an opposite perspective on why so many whites from families with yearly incomes under $6,500 apply to law schools.

I have long felt that the term *admission standards* has been an unfortunate one. It seems to imply the very rat race which Lerner decries. *Requirements* or *prerequisites* would perhaps be more to the point, leaving *standards* to the faculties of our professional schools and the members of our licensing boards, to the competencies necessary to pass courses, graduate, and be certified. A school may insist on prerequisite skills or achievements, but its standards should not be confused with applicants' characteristics.

Our speaker assumes that admissions scores are interpretable in the same way for members of both dominant and nondominant ethnic groups. Some psychologists would generally favor this position (2); others (1) would not. I suspect that there are irrelevant factors in admissions tests for nondominant ethnic applicants which do not obtain for their dominant ethnic competitors. While this issue is subject to scientific resolution and explanation, so far, little has been done to investigate it directly.

Next, as intuitively attractive as Lerner's solution seems at first—to select randomly from among all those who qualify—I don't believe that it is practicable, at least not in the immediate future. For however inadequate the screening instruments may be by themselves or however low empirically determined cutoff scores could be set, I predict that all professional schools will continue to use these tests with undue weight and that the prestigious ones will insist on higher marks than necessary. Given the glut of applicants and a certain desire to save face, would lesser institutions be far behind in their admissions requirements? Also, random selection would encourage students to apply to virtually every professional school in the country, thereby creating problems for institutions and for students, especially those from poverty backgrounds, who too frequently can attend only those schools which are located nearby.

ETS has recently released a study by Evans on the admissions practices in the nation's law schools. Lerner's data on enrollment trends and her contention that there exists a dual system of admissions are supported by this report. According to an article about the Evans study published in ETS *Developments* (4), "The central question addressed . . . was: What would happen if law school admissions committees were forced to disregard racial factors in mak-

ing admissions decisions?" Assuming the continuation of Lerner's "inflated relative standards" by law schools, one analysis of the question, as reported in the ETS publication, estimates that "the decrease in the number of blacks accepted would be 60 percent and the number of Chicanos 40 percent." Only 1 percent of black applicants (10 persons) and 4 percent of Chicano applicants (11 persons) would be admitted annually to the "most selective" law schools. Most blacks and Chicanos—73 percent and 57 percent, respectively—would be found in the "least selective" institutions, whereas only 27 percent of the "White and Unidentified" category would be so situated. If, as Lerner intimates, we were to consider economic criteria instead of race as one basis of affirmative action in admissions practices, the Evans report concludes that this "would not result in the admission of substantial numbers of minority students because the vast majority of low-income candidates is white."

So the conclusion seems inescapable that, constitutional or legal or not, the dual system of admissions has provided access for significant numbers (albeit still not enough) of nondominant ethnic applicants to law schools . . . and to medical and graduate schools as well, including the selective ones.

A Look at the Past

I believe it is important to read into the record of this conference a cursory historical account of the struggle to enhance the access of minorities to graduate and professional schools, an account which our speaker should consider very deliberately.

Dr. Lerner points out that in the 1950s, the number of applications to professional schools necessitated some form of external screening. The fifties also saw a significant rise in minority enrollments, both in undergraduate and in professional schools. Educators involved in equal educational opportunity in higher education have long argued that admissions tests and practices serve not only to keep the number of minority entrants low but also to relegate them to schools with lower standards. The data suggest that this has, in fact, happened. It really should not be surprising, then, to find, as Lerner did, that on the average, minority professional school graduates are earning less than their dominant-group counterparts.

Still, a number of us "blacks" persisted through the fifties and found our way into graduate and professional schools during the sixties. In the meantime, our burgeoning minority populations kept entering undergraduate schools in even greater numbers and percentages. The need to measure the "other relevant abilities" was recognized by us even then, because those of us who gained admission to professional schools (usually after a venture into the world) were aware of how many of our talented, minority ethnic acquaintances were being excluded, not because they weren't competent but because they didn't have the requisite scores. To have one's gifts and talents go undetected is an even deeper frustration than that which Lerner describes.

Many of us have worked hard since those days to ensure that succeeding college graduates have a greater opportunity to attend professional schools than our own peers did, and we have made some gains. We have striven within the field of testing—as yet unsuccessfully—to have measures of these "other relevant variables" developed and to study ethnic differences in admissions test scores; we've been in the courts; we've organized and bargained politically; and we've taken the fight into the streets when necessary . . . and we may again. Only in recent years, when admissions denials have reached embarrassing proportions, have we managed to win a few slots here and there in professional schools. And in rare instances, professional schools have sought us out, have used their own initiative to set aside a few places in order to secure our participation.

I have reviewed these historical developments so that you may understand that all these efforts are not, as Lerner somewhat insouciantly suggests, designed to keep our college graduates out of poverty, but to give our *competent* people the options which are necessary for individual fulfillment and ethnic viability.

Our speaker may argue—and validly so—that these measures have not corrected the basic problems which she has addressed. Nevertheless, we have achieved at least a partial redress, a partial implementation. What is important is that even a small measure of moral justice will not be relinquished to satisfy the Constitution when the result will likely be moral injustice. Lerner has devoted much energy to making her solution compatible with the Constitution. Even now there are developments in the *Bakke* case which suggest that it may be decided on other-than-constitutional grounds (3).

Perhaps our speaker should focus her efforts on demonstrating that her solution is ethnically equitable as well.

Finally, we can live for a while with the discomfort of knowing that preferential admissions will offend many of our citizens, members of dominant and nondominant groups alike. Dr. Barbara Lerner has gotta know that, like Darth Vader of Star Wars we "blacks," we can hang tough!

But we are also profoundly human. So human, in fact, that we, too, can be in touch with the Light Side of the Force.

References

1. Bernal, E. M. Jr. A response to "Educational uses of tests with disadvantaged subjects." *American Psychologist*. 1975, 30, 93–95.
2. Cleary, T. A., *et al.* Educational uses of tests with disadvantaged students. *American Psychologist*. 1975, 30, 15–41.
3. Supreme Court calls for additional briefs in reverse bias case. *Chronicle of Higher Education*, October 25, 1977, p. 14.
4. Supreme Court case sparks admissions bias controversy. ETS *Developments*. 1977, 24, No. 3, 1–2.

Discussion

DEANNE C. SIEMER
General Counsel
Department of Defense

I find the statement of Dr. Lerner's case against dual standards to be persuasive probably because I've argued the same case myself many times in legal briefs and in more informal arguments. I also agree with the assessment of the enormous constitutional cost inherent in any standard based on race. So let me focus my discussion on Dr. Lerner's solution.

I find the statement of the case in favor of a minimum score reflecting competency accompanied by randomization to have several fundamental flaws. I should point out first that these flaws are primarily political and philosophical. This is a solution that, if implemented as Dr. Lerner intends, certainly would withstand constitutional scrutiny and would reduce legal risks substantially. But reducing risks in the courts also involves costs that I think are substantial. Let me try to outline four significant problems.

First, the fundamental assumption of this approach is that such a cut-off score can be set accurately enough to support random selection. If it can't, you simply invite a whole new round of criticism for unjust treatment. The higher the minimum score is, of course, the greater the chance high-scoring candidates will have of being chosen under the random-selection system. If the score is set too low, high-scoring candidates will believe it is unfair to group them in a large category of people which includes many low-scoring candidates because that reduces their chance of being selected. If the score is set too high, low-scoring candidates will have the same arguments they have under the current system. I think the probability that a minimum score can be set accurately enough to support random selection is beguilingly overstated here.

Second, even if a minimum score could be set in a satisfactory manner, this approach assumes that the tests are useful only to predict failure and not to predict relative degrees of success or to maximize success. Several points ought to be made in that regard.

Discussion

It's conceded that test scores correlate with performance in professional school, and the evidence is not at all conclusive that distinctions at the upper end of the scale are not meaningful. Even if fine-tuned distinctions among individuals are not appropriate, this does not mean either that the higher performance predicted by higher test scores is irrelevant or that the use of test scores to maximize group performance is irrational. The paper points out that the score levels now used by law schools for external screening are much too high for judging competency, and the phrases "grossly inflated standards" and "ludicrously inflated standards" are used to describe these score levels. I'm concerned whether there is sufficient evidence that these standards are grossly or ludicrously inflated if relative success rather than probable competence is being predicted. It seems at least politically if not intellectually unsatisfactory to look at minimal scores reflecting competence in a graduate or professional school where enormously valuable and very scarce resources are being used. The use of these resources on candidates who have been selected because they probably won't fail rather than because they will maximize the contribution to be made to the intellectual endeavor seems unlikely to raise enthusiasm in the hearts of the taxpayers or the alumni or anyone else who is contributing to the system. In short, you have to consider whether the public will stand for such a solution. After six months in political office, my instinct is that they won't.

Absent a showing that increased skills do not serve professional competence or that increased skills are indeed negatively correlated with other important character traits, it is entirely appropriate for law schools to try to select the best qualified students by relying in part on test scores to identify those students. In this connection, it seems to me that Lerner's paper confuses the nature of test scores and grades—or at least the results of these two systems. It is not meaningful to describe students as narrowly trying to eke out an additional few points on standardized tests. There is no evidence that tests have an impact on the nature or degree of competition among students. It is the presence of grades that stimulates competition and determines almost entirely the nature of the competition. Moreover, tests play a useful role in correcting for discrepancies in the subjective judgments underlying some grades.

Third, educational institutions have an important interest in avoiding the lowest-common-denominator approach whether that

denominator is phrased in terms of "full competency" or some other classifier. Selecting the most qualified as opposed to the at least minimally qualified students, particularly in a professional school, contributes to the quality of the education offered by the institution. In many situations, the skills of the students affect the pace and the quality of the discussion in the classroom. This is much more than a prestige factor, as described in Dr. Lerner's paper.

Fourth, randomization may be psychometrically pure but it runs strongly against the tradition of merit-based selection which has been and remains a basic tenet of American society. This tradition recognizes the element of gamesmanship in the competition for grades and scores. It recognizes that the A student who gets a 720 on the LSAT may not become a better lawyer than the A− student who gets a 690 on the LSAT. Nonetheless, it's prepared to declare the A student with a 720 score the winner in a competition for a slot at a preferred law school because of the belief that the element of merit is being measured to the extent possible. Only a most determined demonstration that there is no element of merit involved would persuade the public to the contrary.

Randomization also has other drawbacks. Randomization may minimize the feelings of rejection of those who are on the low end of the score totem pole, but you have to take into consideration that it will maximize the feelings of frustration and demoralization of those who are on the high end of that totem pole. On the "feelings scale," if you introduce a system of randomization, you may wind up with a draw.

A Better Solution

In sum, I think a better solution is perhaps the one that Win Manning has discussed extensively in publications and elsewhere: Schools should set minimum standards for acceptable performance in terms of test scores alone and notify candidates that they have the basic ability to succeed in professional school. Then, at a second stage, they should use grades, test scores, scored interviews, scored background questionnaires—whatever means are available—to select from within this competent group those most likely to maximize the contribution to the intellectual endeavor of a graduate school or professional school and beyond that in the profession itself. In so

doing, the chance is minimized that the selection system will undermine public confidence in educational institutions and in the wise use of an enormous amount of the public's money and energy in those institutions.

Session II

Presiding, ROGER T. LENNON
Senior Vice President
Harcourt Brace Jovanovich, Inc.

The Influence of the Law on Professional Measurement Standards

MELVIN R. NOVICK
Professor of Education and Statistics
The University of Iowa

My assignment to speak on the influence of the law on professional measurement standards is a difficult one, as the amount of material that a measurement specialist must master in order to discuss this issue intelligently is awesome. The relevant law must be inferred from a broad corpus of statutes directed primarily to more general issues, from multiple administrative guidelines issued by several federal executive agencies, from conflicting case law, and from a plethora of "definitive" but contradictory legal analyses. More has been written relevant to this issue in the last five years than most nonlawyers could possibly read and assimilate on a part-time basis. Yet, as should become clear as this paper develops, those concerned with educational and psychological testing today cannot do their work without knowing how the law will affect their professional practice. Those of us dedicated to the advancement of the testing profession will need to proceed slowly to try to make some coherent sense out of the legal confusion before us by working closely with legal counsel and then to begin to make such changes in our professional measurement standards and practices as law and circumstances demand. As concerned professionals, we may also decide that it is our responsibility to play an active role in the development of pertinent law.

In my presentation today, I shall not make any predictions as to how various issues will be resolved or when they will be resolved. Rather I shall attempt to identify those activities that we ought to undertake now regardless of how these issues may be settled in the courts or elsewhere. I shall touch on four specific issues and one general issue. First, I shall discuss the question of funding allocation

under Title I of the Elementary and Secondary Education Act (ESEA) of 1965 and indicate how the current discussion about poverty-level versus ability-level funding could affect measurement standards. Then I shall turn to the topic of reverse discrimination as this issue has developed in the context of *Bakke v Regents of the University of California*, a case on which oral argument was heard before the Supreme Court on October 12, 1977 and on which a ruling can be expected this term. Third, I shall discuss the question of federal guidelines for employment selection noting that the first *uniform* guidelines are expected to be released shortly. Then I shall turn to the topic of the reporting of testing irregularities and a discussion of the questions it raises about the relationship between the educational institution and the student as this relationship is mediated by the testing organization. I will also refer to some currently proposed legislation to regulate the testing industry (HR 6776, the Harrington bill). Finally, I will discuss a topic that is central to each of the previous four specific issues: Who is responsible and accountable for educational decision making?

Funding: Poverty vs. Ability

Current funding to school districts under Title I is based on a complicated measure of poverty. The assumption has been that poverty level is a variable that can be measured reasonably well and one that is highly correlated with educational disadvantage. By *educational disadvantage* I mean the net effect of those characteristics of a student's environment that provide less than normal exposure to factors that motivate and facilitate educational growth. Such characteristics affecting educational disadvantage may include, but certainly are not limited to, the following: educational level of the parents, siblings, and peer group; physical facilities in the home; economic status of the family; acculturation of the parents, siblings, and peer group; language proficiency of the parents, siblings, and peer group; quality of formal schooling to date; and possibly, number and spacing of siblings. As a measure of educational disadvantage, poverty level alone has many deficiencies. More complex measures based on a combination of the variables mentioned above will be difficult to fashion, but should in the end prove more valid.

In the 1974 amendment[1] to the Elementary and Secondary Education Act of 1965 mandating the current evaluation of Title I, a new definition of educational disadvantage was introduced for the purpose of evaluation, though no change was made in the formula for funding allocation, which remained entirely a function of poverty level. The new definition reads as follows:

> . . . the term "educationally disadvantaged children" refers to children who are achieving one or more years behind the achievement expected at the appropriate grade level for such children.

This definition would seem either to assume that low attainment accurately indicates the negative influence of the factors mentioned above or, alternatively, to assume that the function of Title I legislation is not to ensure equal educational opportunity but rather to guarantee equal educational attainment. Neither of these assumptions seems valid and, as a result, I think it may be very difficult to present results of the Title I evaluation in an unambiguous way. Professor William Coffman, Director of the Iowa Testing Programs, has pointed out to me that with the Title I evaluation definition of *disadvantage*, the percentage of disadvantaged students in the national norm group for the Mathematics section of the Iowa Tests of Basic Skills rises from 14.5 to 34.25 from the third to the eighth grade. It is not that more students are becoming disadvantaged; only that grade-equivalent scores are being used inappropriately.

The problem becomes more acute with the introduction of HR 7571 (the Quie amendment), a bill that proposes to change the basis of Title I funding to a formula based on ability level. That is to say that Title I money would flow to districts according to the number of students whose test scores revealed them to be, to a specified degree, below the standard for their age. It must be granted that poverty level is not an ideal surrogate for educational disadvantage, and the quantification and scaling of the variables I have mentioned has yet to be done. Yet it might be hoped that a modification of one of these approaches might be found that would be preferable to the proposed ability-level funding. One reason for seeking an alternative approach has been given. Another devastating one follows.

When an educational test is administered, it is assumed that the teachers have motivated their students to do well. When working with disadvantaged minority students, this assumption sometimes is

[1] Section 417 (a) (2) Public Law 93-380, 1974

tenuous, and this is sometimes viewed as a very serious problem. But consider what, hypothetically, would happen if district funding for first grade and beyond depended on students attaining low test scores at age seven. Such a system lends unmistakable incentive for members of the teaching profession, members of local school boards, and taxpayers to arrange for low test scores at this level. This could be accomplished by withholding compensatory treatment at the preschool and kindergarten levels (when it is most effective), by failing to motivate students to perform well on the tests, and by a variety of other strategems. (For a more detailed treatment of this topic, I refer you to the Feldmesser report [4].) If funding by ability level comes into being, we shall need some new measurement standards to help avoid this danger, but it is questionable whether any standards could effectively solve so fundamental a problem.

What, then, is the best available basis for funding? Because of the well-recognized difficulties of poverty-level funding, which have been well documented, the lack of a sound psychometric scaling of educational disadvantage, and the above criticisms of ability-level funding, I would recommend an interim compromise. First, let me say that I think the idea of using the National Assessment of Educational Progress as the basis for census survey data for allocating funds, as studied by Harnischfeger, Huckins, and Wiley (5), is an interesting one. My primary quarrel is with the variable that they were directed to study. What I would recommend is funding on the basis of the distribution of mothers' educational level in each district. Evidence in support of the belief that this is the best available single measure of educational disadvantage is by no means solid; however, further study might support this conjecture. What is important is that mothers' educational level would probably do the required job very effectively with only negligible bothersome side effects and it would be a move in the right direction pending further psychometric development. Certainly the problem of preparing professional measurement standards would be easier if this interim approach were adopted.

The Bakke Case: Measuring Individual Disadvantage

There is no need to review the background of *Bakke v Regents of the University of California*. This has been done in *The New York*

Times, in the *Des Moines Register*, and in the *Iowa City Press Citizen* almost every day for the past two months. What is interesting to me is the position taken by the Justice Department and even more so, the position *almost* taken. While the Department did come out in support of the University of California at Davis and against Bakke, in support of race as an indicator of disadvantage, and in support of goals for affirmative action, the Department stopped short of supporting racial and ethnic quotas. In fact, an early draft of the brief reportedly conceded that quotas are probably unconstitutional.

Just how hard the Supreme Court will come down on quotas is hard to say. Whether they will choose this as the moment to codify the evolution in American political thought away from concepts of group parity and ethnic quotas is uncertain. But that evolution is a reality with or without codification. What this means is that ways other than the use of ethnic or racial quotas must be developed to help eliminate the wide variability in educational advantage now being experienced by young Americans. The only way that I can see this happening is through the quantification and scaling of a measure of *individual* disadvantage with race or ethnicity possibly being used on an interim basis as one of several components of that measure, with the allocation of funding, individual attention, and, in some circumstances, selection being based, in part, on that measure. This same principle can be applied at first-grade level, as indicated previously, and at the professional-school level. Given two applicants with identical Law School Admission Test scores, for example, it should be possible to measure the relative educational advantage each student has enjoyed to that point and to take this into account in the decision-making process, not as a predictor variable but as a component of value or utility. Those who attain a particular standard despite a disadvantaged background are, for a variety of reasons, to be preferred to those from a more advantaged background who attain the same standard. In many cases, such a policy should involve not only preferential selection but also special preparatory treatment before entry into law school. Preparatory schools at all levels, that traditional luxury of wealthy families, ought to be made available, at federal expense, to promising disadvantaged students.

Should the Supreme Court support Bakke in his claim that racial and ethnic quotas are unconstitutional, we shall need a highly ac-

celerated developmental effort for the measurement of disadvantage. We shall need to identify and study all the variables we can find, including racial identification, that relate to disadvantage in a causal way and create one or more measures of disadvantage appropriate for various applications. If the Court fails to rule on the issue, we shall have more time, but the job will still need to be done. Majority tolerance of preferential admission based on racial quotas will not last long nor will minority tolerance of unremedied educational disadvantage. Certainly, variability in educational disadvantage will not disappear overnight, and so we shall indeed need to measure it.

Should the Court choose to take an extreme position and affirm the constitutionality of group-parity concepts and racial quotas, the problems we face will be more difficult. As my colleague, Dorsey D. Ellis Jr., and I (7) argued in the May 1977 issue of the *American Psychologist*, there is no statistically or legally coherent formulation of the concept of group parity and no present standard for racial identification. To implement a policy of racial preference in any scientific or legally acceptable way would require, for example, a definition of *black American* and a measurement procedure for providing racial classifications. To use what for me is offensive terminology, we would, in a reenactment of *Plessy v Ferguson* (1896), be compelled to stipulate what percentage of blackness, say, is necessary for preferential classification, and then we would be compelled to measure percentage of blackness for every person seeking preferential treatment.

Question: Is a person with three white grandparents and one black grandparent himself black or white? What about one black great-grandparent and seven white? And, indeed, were the grandparents themselves black or white?

The support of group parity and racial quotas by the Supreme Court would certainly raise this difficult and distasteful measurement problem. Just that sort of activity is now central to the operation of the Bureau of Indian Affairs, with little benefit to native Americans (see 7).

Let me summarize my position on this issue. To claim that we should be color-blind in academic admissions in contemporary American society is visionary but impractical; to claim that we should have racial quotas to eliminate educational disadvantage is to be ingenuously blinded by color.

Codifying Federal Guidelines

The history of federal guidelines for employment selection is long and complicated. I shall touch only on recent highlights. In 1970, the Equal Employment Opportunity Commission issued stringent guidelines that sought to prevent discrimination against minorities. On November 23, 1976, *Federal Executive Agency Guidelines* were issued jointly by the Justice and Labor Departments and the Civil Service Commission. These guidelines were judged to be consistent with the test standards published jointly by the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education and received much professional support as being a step in the right direction, it being felt that the earlier EEOC *Guidelines* were not consistent with contemporary personnel psychology.

I shall not comment on this topic at any length for two reasons. First, new *uniform* guidelines are anticipated shortly. Second, my discussion of *Bakke* has explored some points relevant here, though there are substantial differences between educational and employment selection. I shall, however, make one point.

It is important to note here that the *Federal Executive Agency Guidelines* on employment selection issued in 1976 refer to the joint test standards and, in fact, defer to these standards for technical detail. Thus, executive agency guidelines effectively having the force of law are adopting professional standards that were not specifically written with this purpose in mind. If, in fact, the new uniform guidelines rely heavily on the joint test standards, and if the uniform guidelines are construed as having the force of law, then a complete review of the testing standards will be in order to determine that they are suitable for this intended purpose. It is entirely appropriate, and indeed highly desirable, that federal regulations defer to professional standards. When this happens, however, we must be sure that the standards can bear this increased responsibility.

The Law and Testing Irregularities

Now I turn to the question of the reporting of irregularities in test administration. This general term covers a multitude of possible kinds of specific events, but the one of primary concern is cheating.

In some situations, proof of cheating is fairly solid. An examiner may see one examinee copying from another or using unauthorized materials and may have the observation confirmed by a second examiner, or handwriting analysis or comparison of fingerprints may establish that an examination has been taken by an impersonator. But in most cases, an allegation of cheating must be confirmed on a probabilistic basis, often through the statistical comparison of responses of two or more examinees. Once such evidence has been evaluated, a mechanism must be available to provide the examinee with appropriate due process in the review of this evaluation and then, if the allegation is confirmed, an appropriate process must be used to notify the concerned institutions of these findings.

Two of the major testing organizations have well-developed policies for the handling of incidents of testing irregularity. These policies are designed to protect the interests of all parties involved. And, indeed, it is in the best interests of all parties, especially examinees, that testing procedures are not compromised. However, policies on irregularities are always written by the testing organization and/or the institution contracting for the testing program. One interested party, the examinee, is not directly represented; his only resource is through the courts. Not surprisingly, there has been some litigation in this area in recent years, though I know of no case in which an examinee has completely had his way in court and prevented a testing organization from withholding his score or cancelling a reported score. However, in *K.D. v. ETS* (6), the New York State Supreme Court (New York's lowest court of general jurisdiction) did insist that in cancelling K.D.'s score, ETS would be required to do so without giving reason for this cancellation. One ruling in one case in one lower-level state court does not establish very much precedent. But it is clear that the content of and manner in which irregularity information is transmitted from a testing organization to an educational institution will be the subject of close legal scrutiny.

In the relationship between the educational institution and the student, the concept of *in loco parentis* is dead but it is not at all clear what has replaced it. Some have argued a trusteeship theory; others have argued for a direct or implied contract relationship. It will probably take time to settle this issue. Moreover, in the testing situation, we have a third party, the testing organization, that contracts with the educational institution to provide a service but

collects the money from the examinees taking its tests. Furthermore, with respect to the contract between the testing organization and the representative of the educational institutions, the negotiation is essentially between equally strong parties. The individual student, however, is powerless. He must either accept whatever terms are established or forego the educational opportunity for which the test serves as gatekeeper. An examinee taking one test administered by Educational Testing Service must sign a statement to the effect that he recognizes that ETS

> [reserves] the right to cancel or withhold any test scores if, in our *sole* opinion, there is adequate reason to question their validity, or if there are grounds for believing that a person has engaged in the act of dishonesty with respect to the testing process. [Emphasis added.]

The question, then, is what recourse does a student have if, hypothetically of course, the testing organization exercises this assumed authority in what the *student* considers to be an arbitrary, capricious, defamatory, libelous, malicious, punitive, or unprofessional way? Well, that is a long complex story, which my colleague, William Buss of the College of Law of the University of Iowa, and I shall tell soon but not today. Lest I create a false impression, let me affirm that there is far more due process in the handling of irregularity cases in the program to which I have alluded than the adversarial nature of the quoted statement would indicate. An ETS example is used here to make a general point because its programs are so highly visible and because ETS management is, as always, so cooperative in providing information.

What I should like to suggest today is that the writing of professional measurement standards for the handling of testing irregularities and the codifying of examinees' rights of due process would be a very useful undertaking. Would it not be desirable for all testing programs in which cheating was a matter of concern to have in their announcements the statement that in the event of an allegation of irregularity, examinees would be afforded due process under procedures consistent with codified professional standards? Perhaps everyone here will agree with this proposed codification of examinee rights. If not, I suggest that we all consider carefully the alternatives:

First, there is the possibility of continuing litigation. While I do not feel that we should solve our measurement problems in the

courts and refuse to predict the outcome of such litigation, I *will* predict that it will continue. In the areas of due process and consumer rights, the public is restless and much less willing to accept even the appearance of arbitrary administrative treatment. De Funis, Bakke, and K.D. challenged the system on important issues, and there will surely be others. You may already have heard the names DiLeo and Clancy. While the next legal challenge to the handling of an irregularity incident is unlikely to receive the attention in the press that the reverse discrimination cases such as *De Funis* and *Bakke* have commanded, I think it will most likely attract more attention than did *K.D.*

Second, if the measurement profession does not adopt standards, there is some chance that Congress may do the job for us. In principle, I am not opposed to federal regulation of industries. But we all know the disadvantgages of federal regulation, and I therefore hope that it would be the response of those involved with educational testing to try to make such regulation unnecessary. Congressman Harrington's bill (HR 6776) to regulate the testing industry is well-motivated. I personally happen to think that a better job can be done by measurement specialists. Furthermore, if this job is done within the profession, we can keep our standards under continuing professional review, modifying or adding to them when necessary. If this job is done by Congress or by delegation from Congress to a federal agency, the measurement profession may have as little influence on these standards as they have until recently had on employment selection guidelines. If we must have federal legislation, we ought to have the kind that defers to professional standards on all points of substance.

Protecting All Participants in the Testing Process

There are many issues relevant to the four topics I have discussed that will require codification in professional standards, but in my thinking, one predominates. There are three parties involved in the testing process—the examinee, the educational institution, and the testing organization. It has, for example, been alleged that intelligence testing has been used to track minority children in such a way as to limit their intellectual development. If this is true, some action should be taken. But who is responsible? Is it the school

which claims that these procedures are justified by the scientific evidence in support of the validity of the test? Or is it the test publisher who acknowledges the responsibility to produce good tests but professes to lack responsibility for specific uses of the tests and the specific decisions which are based on test scores? To what extent can and should test publishers be required to monitor the use of their tests or the scores they report? Should they be able to say simply that they only publish or give tests and that they have no responsibility for the decisions that are made? Alternatively, can testing organizations institute mandatory licensing procedures that permit them to control the use of their tests, or is this restraint of trade and/or infringement on individual professional practice?

I suggest that these questions be answered only by the creation of legally and scientifically valid professional measurement standards that respect and codify the rights of all participants in the testing process. It is my judgment that educational testing has been a major force in the democratization of the educational process. The codification of new standards in the areas I have mentioned will, I believe, further strengthen this force.

References

1. *Bakke v Regents of the University of California*, 18 Cal. 3d34, 132 Cal. Rptr. 680, 553 P2d 1152 (1976), pending U S Supreme Court (1977).
2. Educational and Employment Opportunity Commission guidelines. *Federal Register*, November 24, 1976.
3. Federal Executive Agency guidelines on employee selection procedures. *Federal Register*, November 23, 1976.
4. Feldmesser, R. A. *The use of test scores as a basis for allocating educational resources. A synthesis and interpretation of knowledge and experience*. Princeton, N.J. Educational Testing Service, November 1975.
5. Harnischfeger, A., Huckins, I. F., & Wiley, D. E. *The National Assessment of Educational Progress model. A tool for achievement-based Title I fund allocations*. Chicago: MP-Group for Policy Studies in Education, CEMREL, 1977.
6. *K. D. v Educational Testing Service*, 87 Misc2d 665, Supreme Court, Special Term, N.Y. Court, July 26, 1976

7. Novick, M. R., & Ellis, D. D. Jr. Equal opportunity in educational and employment selection. *American Psychologist*, 1977, 32, No. 5, 306–320.

8. *Plessy v Ferguson*, 163 U.S. 537 (1896).

9. *Standards for educational and psychological tests*. Washington, D.C.: American Psychological Association, 1974.

10. Title I: Financial assistance to local educational agencies for the education of children from low-income families. Elementary and Secondary Education Act, 1965, as amended by Public Law 93-380 (August 21, 1974) Section 417 (a)(2).

Discussion

WINTON H. MANNING
Senior Vice President for Development and Research
Educational Testing Service

When I was a youngster growing up in St. Louis on the banks of the Mississippi—not so terribly far from Mel Novick's Iowa domain—a favorite pastime was scaling stones. A boy or girl who could get three skips was a star; if you could get four hops before the final splash, you were a superstar. Mel Novick would probably have earned the equivalent of an Olympic medal because he has taken careful aim and with characteristic vigor and style, scaled five current and crucial areas in which testing and the law intersect.

His five skips are:

1. funding allocations under Title I
2. admissions policy and the Bakke case
3. guidelines for use of tests in employment selection and licensing
4. fairness and due process in the treatment of students suspected of cheating or similar irregularities
5. the general question of the allocation of responsibility (and accountability) in educational decision making, which is implicit in all the foregoing.

(The last is the final splash, and authorities differ on whether that counts or not in the Olympic point count for the stone-scaling event.)

The problem I now confront is how to discuss adequately Novick's *tour de force* in 10 minutes or so. Upon reflection, I decided the only answer I could give, if asked whether I could do it, is the one that a fellow St. Louis lad, Yogi Berra, gave when asked if the Dodgers would sweep the World Series. Yogi replied, "I can answer that question in two words—IM POSSIBLE!"

I have now used up two minutes establishing that I am a human being with warm memories of childhood idylls and that you, as a

sensitive and intelligent audience, should sympathize with me in my predicament. Nowadays these are not mere ritualistic rites of platform palaver because, as I think Roger Lennon would agree, any officer of a large testing organization should probably assume that a substantial number of people in any audience he faces will believe automatically that he is *not human*, never had a *childhood*, and deserves all the *hostility* he gets.

I shall now scale my stone:

1. Title I

Mel has pointed out some pitfalls in implementing an achievement test-based formula for allocating Title I funds among the states and school districts. The income-based formula for Title I allocations is an exceedingly complex formula. For example, it begins with the poverty-line basis for determining eligibility for federal programs which, in turn, rests upon the 1970 census data on state-by-state income level, size of family, and sex of head of family, and data on the number of 5-to-17-year-olds in a family. On this foundation, special adjustments are introduced to take into account:

1. A percentage of the AFDC children (that is, *A*ffluent *F*amilies with *D*ependent *C*hildren, the term applied to children of non-poverty or "wealthier" families);
2. Some percentage of the children not in families but in non-state-operated placements, such as foster homes, private institutions, and so forth;
3. Some percentage of the handicapped and migrant children in the state;
4. Annual updating to reflect changes in the Consumer Price Index;
5. And finally, adjustments reflecting the individual state's average per-pupil expenditure for recent years (specifically what it was three years ago), the national average per-pupil expenditure, and past funding levels within the state. All of these are also subject to change and demand frequent updating.

Mel has rightfully pointed out that a move to an achievement test-based eligibility formula raises a number of knotty measurement problems, but I am not sure that the present income-based model is,

on balance, any more satisfactory. In summary, I am personally not as pessimistic as Mel is concerning the feasibility of allocating funds by tests. The best service I can offer to you who are interested in this question is to commend for your reading an excellent NIE study just completed which is entitled "Using Achievement Test Scores to Allocate Title I Funds."

The decision regarding which basis to use—income or ability—should rest upon value considerations rather than technicalities of measurement. Title I was originally conceived to benefit poor people by breaking the cycle of poverty through educational intervention. Congress has drifted to a confused middle ground in more recent years. If they wish to direct funds to educational need *per se*, then achievement test-based allocation systems make a great deal of sense to me. If poverty is the central concern, income-based eligibility seems relevant. It is a matter of social policy on which Congress badly needs to clarify its intent.

2. Federal Guidelines for Employment Testing

I cannot really find anything with which to disagree in Mel's discussion of this matter, and I know that Wayne Holtzman will discuss this at greater length this afternoon. So I will pass over this quickly. Furthermore, I have recently presented a lengthy paper on just this topic, and once started on these issues, it would be hard for me to stop. In my AERA paper (entitled "Educational Research, Test Validity and Court Decisions"), I said that a decade of experience with the APA *Standards* and the EEOC *Guidelines* suggests that there is a serious need for reconceptualizing the theory of test validation implied in these two documents. In fact, I offered the same judgment as that visited upon the Emperor Galba by Tacitus:

"*Omnium consensus capax imperii nisi imperasset.*", which means in plain English:

"Had he never been placed in authority, nobody would ever have doubted his capacity for it."

I hope I may be pardoned this pretentious allusion to the classic Latin, but when one is surrounded by lawyers, it is well to remind them that even psychometricians can use dead languages to make their point!

3. Fairness and Due Process in the Handling of Suspected Cheating Cases

There are few more uncomfortable, unsatisfying, and complicated topics than this one. Mel was right to raise it, but I don't think he did it justice, as I hope he would agree. I look forward to reading the paper he has promised us on this matter. I would simply add some points that he overlooked, in order to expand the context of the discussion:

- At ETS, no student tested in the United States ever has his scores cancelled, regardless of how extensive the evidence of alleged cheating, without first being offered the opportunity to take the test again. If he or she does agree to be retested and the prior test performance is confirmed, the matter is closed. It is also relevant to add that the Appellate Division Court in New York (a higher court than that which Mel mentioned in the *K.D.* case) has commented favorably on the practice of offering a retest, stating that spending two and a half hours taking the test again seems not to be an undue burden on the student. (Incidentally, it is probably a lot more generous treatment than is typically given to students by many professors, although I am not accurately informed about practices at the University of Iowa in like circumstances.)

- ETS has set forth in one document a clear statement of the principles and procedures we follow in cases of irregularity—the due-process considerations, if you will—and this document is provided to every candidate who finds himself in a situation where his scores are questioned.

- In the Law School Admission Test program, a procedure involving arbitration is employed, thus reducing time and expense for the candidate that a court case might require. The law school program is, by the way, the only ETS program in which a reason is given for score cancellation, a policy adopted by the Law School Admission Council after careful deliberation. I realize this does not address all the issues Mel has raised, but it is important to dispel, to a degree, the impression that his understandably brief comments may have left with many of you.

4. Fairness in Admissions, as Exemplified in the Bakke Case

This is by far the most important and complex issue that Mel has addressed. As one of the authors of the recent Carnegie Council report on the subject of race in admissions to higher education, it is necessary that I first set forth clearly where I come from on this matter before turning to the more provocative elements of Mel's discussion.

The central social and educational issue of the *Bakke* case is the problem of balancing considerations of individual and group equity, a problem which turns upon difficult value choices. Not all individuals and not all institutions will agree about them. In this circumstance, the public must have *confidence* in the process by which decisions are made. The selective professional schools in particular must be prepared to face public scrutiny of their processes and their policies; and both the processes and the policies should conform to their own missions and to the demands of public policy and should be fair, as among individuals similarly situated. Above all, these schools must be concerned with making optimal use of their facilities to develop human resources for service to society. In the effort to reach this goal, *racial experience is relevant* within the admissions process because important educational and professional objectives will not be attainable unless, as colleges and universities go about the task of making admissions decisions, *consideration is given to the minority status of individual applicants*.

Mel has suggested that the Supreme Court's decision may require that ways other than the use of ethnic or racial quotas must be developed to help eliminate the wide disparities in educational advantage of young Americans. He then goes on to say, and I quote, "The only way I can see this happening is through the quantification and scaling of a measure of *individual* disadvantage, and the allocation of funding and *individual attention* being based on this measure." (Emphasis supplied.)[1]

I have no quarrel with the principle of allocating funds on the basis of an aggregated measure of disadvantaged status, or even its consideration within a context of examining a particular candidate's

[1] **Editor's note:** The sentence quoted above is from the draft of Mr. Novick's paper that Mr. Manning used as a basis for his discussion. The sentence was revised in the final version that appears in these *Proceedings*.

background. I question, however, the desirability of making individual admissions decisions about particular students on the basis of such a standardized test of disadvantaged status.

It *is* an intriguing idea, one that I have played around with a bit myself. In fact, I constructed such an index of disadvantaged status, based upon family income, parental education, and whether English is spoken in the home or not, and then applied it to data on 16,000 students (4,000 blacks, 4,000 whites, 4,000 orientals, and 4,000 Hispanic students). There is a modest increase in the proportions of minority students found in the upper 10 percent of the distribution of test scores, after these factors are regressed out, but *not much*. I don't want to belabor this point; you can read about it in my Carnegie Council report (4). Frank Evans, using a similar logic, got the same disappointing results using law school admissions data (3). If you use socially disadvantaged status rather than racial experience, the proportions of minorities selected will be substantially reduced over present levels; disadvantaged status is no substitute for taking race or ethnicity directly into account if you are concerned about maintaining enrollment levels of minority students.

But his is not the only point we should be debating, as I am sure Mel would agree. The point is whether giving each child a score on a scale of disadvantagement is good social policy. Does it lead to the sort of society we want? Would you want to establish a quota, for example, for the number of students scoring below 500 on the Educational Disadvantagement Scale?

In the effort to strike a balance between the rights of groups and the rights of individuals—between "a pluralist society and a nation of individuals"—I fear that the proposed instrument is too frail a bridge to bear the traffic. As an instrument for research analysis, as a means of dissecting the social consequences of aggregations of individual admissions decisions, I can find no quarrel with the Novick proposal. But I would part company with Mel if such a scale were to become the explicit basis for admitting and rejecting students. A writer in *The New Republic* (2) recently put it well:

> To be classified is to be judged as a member, not as a particular person. That is doubly threatening first as to individual rights and then to the integrity of community life. Classification for a purpose other than sheer description [the analytic function to which I have alluded] implies a hierarchy of classes. So it drives people to choose their groups, if they can, for reasons other than their private feelings and

commitments . . . Admissions officers should look for personal strength—pride, energy, enterprise, compassion—with the understanding that these qualities are differently expressed in different cultures, and tested far more harshly in some parts of our society than in others. But personal strength, by definition, is an individual trait, not a group trait. It cannot be recognized unless all groups are treated as individuals. It is difficult to do that with any degree of fairness in an egalitarian society. *But that is what doing justice requires.* [Italics supplied.]

Simple justice, I believe, requires that admission officers take racial experience—particularly evidence of having surmounted barriers of racial discrimination—into account as they look in depth at each applicant. The explicit use of scales of disadvantaged status as a part of a student's credentials should be approached with great caution.

We Must Proceed Slowly

Twenty-two years ago, Professor Edmund Cahn, writing in the *Annual Survey of American Law*, observed that "we should not make our constitutional liberties a function of anybody's science." Law remains free because the principle upon which it rests is an ideal; measurement, broadly conceived, is no more than the systematic quest for regularities in human behavior (1). Although it is inevitable and necessary that law and social science consort with one another, their interaction is fraught with peril to each.

I wish we had a whole day to talk about Mel Novick's provocative paper. We might even skip a few stones as we follow his own injunction—namely, that "those of us dedicated to the advancement of the testing profession will need to proceed slowly to try to make some coherent sense out of the legal confusion . . ."

References

1. Coons, J. L. Recent trends in science fiction. Serrano among the people of numbers. *Journal of Law and Education*. 1977. 6. No. 1. 23–40.
2. Disadvantaged groups, individual rights. *The New Republic*, October 15, 1977. 5–9.
3. Evans, F. R. *Applications and admissions to ABA accredited law schools. An analysis of national data for the class entering in the fall of 1976.* Report No. LSAC-77-1. Princeton, N.J.: Educational Testing Service, 1977. 1–108.

4. Manning, W. H. The pursuit of fairness in admissions. In *Selective admissions in higher education*, a report of the Carnegie Council on Policy Studies in Higher Education. San Francisco: Jossey-Bass, 1977.

Editor's Note: Thaddeus Holt, Attorney, Washington, D.C., also discussed Dr Novick's paper.

Session III

Presiding, ALFRED B. FITT
General Counsel
Congressional Budget Office

Validity and Legality

WAYNE H. HOLTZMAN
President, Hogg Foundation for Mental Health and
Hogg Professor of Psychology and Education
The University of Texas at Austin

For several generations, psychologists and other specialists in educational measurement have struggled with the concept of test validity and how best to determine it. The basic idea is deceptively simple. When we raise a question about a test's validity, we want to know whether the test really measures what it purports to measure. How valid is the test for a particular decision or interpretation proposed for it? No matter how excellent a test is in other respects, if it measures the wrong thing or is wrongly interpreted, the test is worthless. Our chief concern here is with the use of tests and educational measures for making decisions that affect individuals. The kinds of questions ordinarily asked are the following: How valid is this test for deciding whether or not an individual is qualified to enter law school? Can I use this test for making a decision as to whether or not an eight-year-old child should be placed in a special education class for slow learners? Is this test a good one for selecting the best applicant for a limited number of job openings in my company? The more valid a test is for a particular selection or classification procedure, the more likely one can avoid making mistakes in using it as a basis for decisions about individuals.

The rapid growth of testing as a means of personnel selection and classification during World War II clearly demonstrated that even tests with relatively low validity can be of great practical use if one is able to select only a small number of individuals from a large number of applicants. If one selects only 50 individuals for flight training out of 500 who wish to become pilots, the important consideration is to reject all the doubtful cases, taking only the cream of the crop. When the cost of training an Air Force pilot runs upwards of $100,000 per pilot, there is obvious social value in using very high selection ratios with extensive batteries of tests to insure that the ones admitted to the training program are the individuals most likely to succeed. This classical model of selection and classification works

fairly well as long as one is not concerned about what happens to the very large numbers who are rejected in the process. The fact that many rejected applicants might indeed have been outstanding successes if they had been selected is irrelevant as long as those who are accepted do succeed.

During the 1950s and early 1960s, growing awareness of social injustice in the United States led to sweeping new legislation such as the 1964 Civil Rights Act. Title VII of this act was designed to achieve equality of employment opportunities and to eliminate discrimination in hiring and promotion practices. The act was a strong one leading to numerous challenges of employment practices, many of which involved the use of tests for selection and classification of personnel. Discriminatory purpose on the part of an employer need not be proved, and it is insufficient merely to demonstrate a rational basis for the challenged practice. Shortly thereafter, the Equal Employment Opportunity Commission developed its *Guidelines on Employment Selection*. The *Guidelines* leaned heavily upon the 1966 edition of the American Psychological Association's *Standards for Educational and Psychological Tests and Manuals*, a set of professional test standards developed jointly by the major professional organizations concerned with educational and psychological testing.

The EEOC *Guidelines* placed a heavy burden upon the test user to prove that there was sufficient predictive validity in the test, when used for hiring or promoting minorities as well as whites, to meet the requirements of nondiscriminatory selection. Other approaches to validation were generally set aside as insufficient. This strong posture soon led to a series of law suits that have recently culminated in several important Supreme Court rulings. Not only Title VII of the 1964 Civil Rights Act but also the Equal Protection Clause of the Fourteenth Amendment to the Constitution has played an important role in these Court decisions. Before examining these decisions in more detail as they relate to questions of validity, let's take a closer look at the ways in which test validity is generally established.

Establishing Validity

Determining the validity of a testing procedure can be done in several different ways. If the procedure is aimed at selecting indi-

viduals who will be successful on the job or in school at some future date, then the ideal method is to study a large number of individuals who have been given the test prior to admission to see who the successes and failures are at a reasonable later point in time. Correlation between the test and the criterion yields a coefficient of *predictive validity*. A second approach often considered a variation of predictive validity involves obtaining both test scores and criterion measures on the same sample at the same time. If a new test correlates highly with an established valid test when given to the same subjects at the same time, *concurrent validity* is established for the new test using the previously validated test as a criterion measure. A third approach, known as *content validity*, is particularly applicable when no criterion measure is available and the focus of concern is upon measuring certain knowledge or ability, the relevance of which is self-evident. A fourth method of establishing validity involves relating the test scores to a theoretical concept or construct by conducting a series of interlocking experiments designed to test hypotheses based upon theory. Such *construct validity* is only slowly established as hypotheses are confirmed, refuted, or revised. All four of these methods are well established as fully acceptable procedures, but one must be careful to fit the method to the particular purpose he has in mind.

Establishing Predictive Validity

For selection procedures in which individuals are admitted or rejected on the basis of a test score, establishing predictive validity is of primary concern. There are four conditions that must be met in order to establish predictive validity for a decision rule based upon test scores. First, one must have a good criterion against which performance on the test can be compared. If the purpose of the test is to select individuals who will succeed in graduating from college, one must wait at least four years from the point of admission until the criterion can be obtained. If the test is to be used for selecting the best applicants for a job, obtaining adequate criterion measures may be even more difficult, depending upon what is most highly valued in the way of job performance. In most cases, there is no single criterion measure that is satisfactory. Criterion measures may be just as multi-dimensional as the tests used to predict them. A test

may prove to be highly valid for predicting one criterion and quite unsatisfactory for predicing another within the same selection procedure. If one insists that a test must predict later success in a career rather than merely success in completing a training program, a delay of many years may have to occur before both test performance and criterion performance are measured in order to validate the test.

A second condition of importance in establishing predictive validity is the admission of a full range of test scorers, both high and low, for whom criterion data are obtained at a later date. When a test is used for selection so that only the top applicants are admitted, there is no way of determining later whether those who were rejected would have been failures. In most practical situations, it is very difficult to maintain sufficient heterogeneity for an adequate determination of predictive validity.

Where serious questions of bias or unfairness arise in a testing procedure, as in the case of a test which may be judged biased against blacks in favor of whites, one must have a sufficient number and heterogeneity of the disadvantaged individuals as well as those who are favored in order to establish separate validity coefficients for both classes of individuals. This third condition, vigorously pursued by the Equal Employment Opportunities Commission, is particularly difficult to meet in most practical situations. In industries where blacks have been discriminated against until recently, there is an insufficient amount of evidence from previous research to defend the test procedure against charges of bias.

A fourth condition for establishing predictive validity in a selection procedure follows from the requirement that a good criterion measure must be obtained at a later date against which the test scores can be compared. The lapse of time between test scores and criterion data must be sufficient to allow for the development of an adequate criterion measure. During this period, which may range from months to years, there may be considerable attrition due to dropouts or turnover in personnel. If this attrition rate is appreciably different for whites than it is for blacks or Chicanos, the outcome of the validity study is considerably weakened.

For all of the above reasons, establishing predictive validity is exceedingly difficult and, indeed, impossible in many cases where tests are used. The best studies are usually those carried out under the auspices of major social institutions with thousands of individuals

participating, such as in military or educational settings. But more often than not, the sample sizes are too small to determine the practical significance of the selection procedure; there are too few cross-validation studies to determine how strong the validity is under varying conditions; there are too few disadvantaged individuals in the validation study because of a previous history of discriminatory employment practices; and weak or irrelevant criteria are employed. Given this state of affairs, it is small wonder that vigorous enforcement of Title VII in the 1964 Civil Rights Act, as interpreted by the Equal Employment Opportunities Commission, has resulted in a number of federal court cases involving the use of tests in personnel selection and classification. Three of these employment test cases have been carried to the Supreme Court: *Griggs v. Duke Power Company*[1] in 1971, *Albemarle Paper Company v. Moody*[2] in 1975, and *Washington v. Davis*[3] in 1976. It is instructive to examine briefly each of these cases with respect to their implications for test validity and legality.

Court Rulings on Tests

In *Griggs*, black laborers at the Duke Power Company's generating facility filed suit against the company, challenging the company's requirement that an employee must have a high school diploma or must pass an intelligence test in order to be transferred from the Labor Department into one of the other four departments at the power plant. Since blacks were employed only in the Labor Department where the highest paying jobs had lower wages than the lowest paying jobs in the other four departments, it is clear why Title VII of the Civil Rights Act was used as a basis for this lawsuit. The Supreme Court ruled in favor of the Negro employees. The Civil Rights Act requires that a selection test which discriminates on the basis of race is prohibited unless the test can be shown to be clearly related to job performance. While the burden of proof for showing discriminatory consequences rests upon the plaintiff, once such proof has been presented, the burden shifts to the defendant who

[1] 401 U S 424 (1971)

[2] 422 U S 405 (1975)

[3] 426 U S 229 (1976)

must prove that the test does indeed correlate with job performance. As Lerner (1) has pointed out in a review of these employment testing cases, the lower courts quickly applied the *Griggs* definitions to cases arising out of very different contexts. The net result was considerable distress and confusion over the differing bases for proving discriminatory effects on the one hand and for proving test validity with respect to job performance on the other. While the Supreme Court ruling four years later on the *Albemarle* case provided some clarification, it wasn't until *Washington v. Davis* that significant new features were considered, leading to at least a partial resolution of the problems raised by these lower court cases following *Griggs*.

The action in *Washington v. Davis* started in 1970 when two black police officers filed suit against the Commissioner in the District of Columbia alleging that the promotion policies of the police department were racially discriminatory. The department's recruiting procedures involved a written personnel test that excluded a disproportionately high number of black applicants. The test had been developed by the Civil Service Commission and was designed to measure verbal ability, vocabulary, reading, and comprehension. The district court upheld the use of the test by the police department, but the appeals court reversed the decision, ruling that the test had not been established as "job related." In 1976, the Supreme Court reversed the appeals court, ruling in favor of the police department and its use of the test for selection purposes.

The Supreme Court's written opinion in *Washington v. Davis* is of particular interest because it includes detailed reference to test validity. The Court noted that there is no single method for appropriately validating employment tests in relation to job performance. A validation study of the test had been completed by the Civil Service Commission, which demonstrated a relationship between test score and performance in the recruit training program for police officers. The Supreme Court joined the earlier district court in concluding that this evidence was sufficient under the Equal Protection Clause of the Fourteenth Amendment. One month later, the Supreme Court extended these concepts further by denying plaintiff's petition for review of *Tyler v. Vickery*, a case concerning the use and validation of the Georgia Bar Examination which the plaintiffs claimed was racially biased and invalid. As Manning (2) pointed out in a recent analysis of test validity and court decisions, the Fifth

Circuit Court of Appeals in *Tyler v. Vickery* revealed a new disposition of the courts regarding what constitutes evidence of test validity in cases in which a disproportionate number of black applicants, as contrasted to whites, fail an examination. The petitioners who had been denied admission to the bar by the test argued that the bar examiners could justify the use of their test only if they were able to show a predictive validity correlating test scores with later occupational performance. Expert witnesses for the bar examiners differed sharply on the issue of test validity, claiming that the bar examination would be sufficient if it had content validity. The court rejected the arguments of the petitioners—namely, that the user of a test with discriminatory results must prove the job-relatedness of the test—and held that the Equal Protection Clause of the Constitution requires nothing more than that an examination test skills and knowledge which have a logical, apparent relationship to those necessary for the job. For the first time, contrary to the *Guidelines of the Equal Employment Opportunities Commission*, the courts rejected exclusive reliance on a criterion-related approach to the proof of job-relatedness, ruling that the proof of content validity would be sufficient.

An even clearer case supporting content validity is the *United States v. South Carolina*,[4] as decided by a three-judge district court in 1977. The South Carolina State Board of Education had been using the National Teacher Examinations to certify and determine the pay levels of teachers in South Carolina. The test had been used in South Carolina for over 30 years, and a considerable amount of test data had accumulated on both black and white teacher performance. Extensive validation studies had been undertaken by Educational Testing Service which resulted in different minimum test scores in various areas of teaching specialization. The plaintiffs challenged each of these uses of the National Teacher Examinations, contending that more blacks than whites historically had failed to achieve the required minimum score. In its opinion, the court cited in detail the design of the validation study and ruled that it was sufficient to meet the burden placed on defendants under Title VII of the Civil Rights Act. The study demonstrated content validity by measuring the degree to which the content of the tests matched the content of

[4]In the District Court of the United States for the District of South Carolina, Columbia Division, Civil Action No 76-1610, filed April 14, 1977

the teacher-training programs in South Carolina. A minimum score requirement in each of the major fields of specialization was established by estimating the amount of knowledge that a minimally qualified teacher candidate would have in South Carolina. A total of 456 teacher-educators in South Carolina from both races participated in the content validation of the National Teacher Examinations. The arguments presented go further than any other court opinion in supporting content validation as sufficient, even in the face of unintended discrimination. How *United States v. South Carolina* will fare when and if it is appealed remains to be seen. Regardless of final outcome, conflicts between the EEOC *Guidelines* for test validation and rulings by the federal courts have increased sharply in the past 18 months since *Washington v. Davis*.

What do these trends of the past six years suggest with respect to test validity and legality? Clearly there is less burden now upon the test user to establish the predictive validity of the test procedure than there was several years ago when the lower courts were extending *Griggs* into unknown territory and when the EEOC *Guidelines* for test validation prevailed. For the reasons cited earlier, good predictive validation studies are difficult, if not impossible, to achieve in many cases. The use of intermediate criteria in training situations rather than job performance, the adoption of multi-dimensional criteria as well as multi-dimensional assessment procedures where feasible, and a heavier emphasis upon content validation when dealing with specific knowledge or skills can provide a stronger basis for accepting or rejecting a testing procedure than the deceptively simple unidimensional prediction of a questionable criterion. As Messick (3) has pointed out in a discussion of meaning and values in measurement, advances will be made only by adopting a more comprehensive framework than either predictive, concurrent, or content validity can provide when considered alone. For Messick, all measurement should be linked to constructs which serve to organize disparate findings. Proof of validity must rest upon both empirical and logical analysis if it is to prove sufficiently robust to withstand challenge in the courts.

Conclusions

In conclusion, what can be said at this time concerning the relationship between test validity and legality? In my opinion, the following

points are worth noting:

1. A testing procedure that incidentally discriminates against one or more classes of individuals can only be justified if job-relatedness is demonstrated. Once proof of discrimination has been demonstrated by a plaintiff, the full burden of proof of test validity rests with the test user as defendant. The test user does not have to prove the job-relatedness of the test by a predictive or concurrent validation; a logical, apparent relationship between the test skills or knowledge and the job by content validation is sufficient.

2. An excessively heavy burden upon the test user to prove job-relatedness by predictive validation has been lightened substantially. The burden of proof is present, but a greater choice of alternative validation procedures has been recognized by the courts.

3. Predictive validation of a testing procedure for personnel selection is still the method of choice when a good criterion measure can be obtained, when adequate samples of appropriately heterogeneous individuals are available, and when differential attrition due to extraneous factors is not markedly different for identifiable classes of individuals such as whites, blacks, and Chicanos.

4. Immediate or intermediate criteria can be a satisfactory substitute for ultimate job-related criteria in concurrent or predictive validation of a testing procedure, provided the relevance of the substitute criteria can be demonstrated.

5. Content validation is a fully acceptable procedure for personnel classification using tests of knowledge or skill (a) where the samples of items or tasks to be performed can be demonstrated as reliably representative of the specific domain of the abilities in question, (b) where there is expert consensus that the domain is relevant to the training program or job requirements, and (c) where irrelevant difficulties have been eliminated. An example of an irrelevant difficulty is the inadvertent requirement of unduly high verbal ability to understand items in a written test of mechanical ability for selection of auto mechanics.

Undoubtedly, additional court decisions will be necessary to clarify further the complex and evolving relationships between the validity and legality of selection and classification procedures involving tests. Decisions about human beings are clearly too important to be left primarily in the hands of psychologists or educators, regardless of advances in psychometric theory and practice. It is incumbent upon all of us to listen carefully to what the courts have

to say. Recent decisions and their supporting opinions have once again broadened the range of recognized options for validating test procedures. At the same time, the limitations of traditional tests with their heavy emphasis upon verbal skills have become increasingly apparent. Only by the accelerated development and validation of new techniques for measuring other important personal characteristics can we hope to achieve both social justice for the disadvantaged and equal protection for all individuals.

References

1. Lerner, B. Washington v. Davis: Quantity, quality and equality in employment testing. In Philip Kurland (Ed.). *Supreme Court Review* (1976). Chicago: University of Chicago Press, 1977. Pp. 263–316.
2. Manning, W. H. Educational research, test validity and court decisions. Unpublished manuscript, April 1977.
3. Messick, S. The standard problem. Meaning and values in measurement and evaluation. *American Psychologist*, 1975, 30, 955–966.

Some Possible Social Implications of Recent Court Decisions

Charles L. Thomas
Program Coordinator
Minority Affairs Program
Research Triangle Institute Center[1]

Two considerations make any intelligent discussion of the social implications of recent court decisions most difficult and perhaps speculative: For one, the possible resolution of the conflict between equal access to educational opportunities and the protection of individual rights is still pending in the U.S. Supreme Court (*Bakke v. the Regents of the University of California*). Future consequences, to a great extent, will depend upon how the Court decides and whether the decisions should be given a narrow or broad interpretation. Second, one must realize that any efforts to identify all or even a substantial portion of the malicious or beneficial consequences of perhaps the most critical decision before the court since *Brown* will surely fall short of its mark.

Thus, armed with these caveats, I now will attempt to venture into these murky waters, cognizant that the rocks underneath on which I must tread make for slippery footing. However, I have taken the precaution of interjecting the words "some possible" in front of the title of this presentation.

One strategy for determining the possible consequences of recent litigation involving access to opportunities for minorities is to extract the common elements reflected in the concerns of the more vocal opponents and proponents of affirmative action programs in general and special collegiate admission programs in particular.

[1]The author, who is currently on leave from Indiana University, assumes full responsibility for all views expressed herein none of which should be construed as reflecting the policies or beliefs of RTI or Indiana University

Another approach is to focus, by logical analysis, on the specific impact that an unfavorable decision by the Supreme Court may have on various elements of affirmative action. Admittedly, the latter approach rests on the assumption that the idea of affirmative action in principle is neither repugnant nor unconstitutional.

Before addressing these concerns, it may be best 1) to turn briefly to the "compelling reasons" for the past legislative responses designed to increase the likelihood of opportunities for minorities (the "compelling reason" principle, as we shall see, has played a crucial role in recent court decisions); 2) to outline the more salient federal laws instituted to meet these needs; and 3) to outline the trends in litigation that have emerged because of either voluntary attempts to meet the social need and the spirit of the law or because of policies and practices that have intentionally or unintentionally circumvented the law.

The "Compelling Reasons"

The devastating effects of slavery and past and present racial and sex discrimination have been documented and rarely denied. Nor has the generation-to-generation link between poverty, despair, and educational and psychological disadvantages among racial minorities been refuted.

To be sure, ethnic minorities such as blacks, American Indians, and Hispanics, as well as women, have made significant advances in recent decades. For example, statistics on white collar employment show gains for nonwhites and Hispanic Americans during the decade of the 1960s, unemployment dropped from approximately 11 percent in 1961 to less than 5 percent in 1969 for nonwhites 20 years of age and over; and the rate of participation by women in the labor force increased from 33.3 percent in 1950 to 44.7 percent in 1973 (21).

However, in spite of the gains, the overall employment picture for minorities and women remains discouraging. In terms of the percentage of the civilian labor force that is unemployed, little has changed: The nonwhite, white ratio of unemployment was 2:1 in 1973 as it was in 1954 (25); the median income of blacks stood at 58 percent of the white median income in 1973, the same as found in

1954[2] (22); women's 1973 earnings were only 58–60 percent that of men's income (23); and women and minorities continue to be underrepresented in the high-paying prestige jobs and overrepresented in employment requiring fewer skills and, of course, paying the lowest wages (6); precipitated by the 1974 recession, the jobless rates reached 14.3 percent for nonwhites and 12.4 percent for Spanish Americans in 1975 and then peaked in 1976 at 14.5 percent for nonwhites.

However, the most disturbing and incendiary statistics are the jobless rates among minority teenagers of 16–19 years old. In 1954, the unemployment rates for nonwhite male and female adolescents were 14.4 and 20.6 percent, respectively; by 1964, the rates had climbed to 24.3 and 31.3, respectively; and by the first quarter of 1975, black teenage unemployment had climbed to 39.8 percent compared with 18 percent for white teenagers. (Unemployment for white teenagers has been rising steadily, ranging from approximately 12 percent in 1954 to 14.1 percent during the third quarter of 1974 [17].) In the large urban centers, the estimated unemployment for minorities, particularly blacks, may be as high as 50–60 percent.

These statistics are alarming for several reasons: 1) A large number of the 16–19-year-olds, though closer than their older counterparts to the usual point of entry to higher education or job training, are receiving neither. 2) the lack of training and employment at these young ages will likely assure the cyclical pattern of poverty that typically characterizes the urban and rural minority environments; 3) the feeling of hopelessness about obtaining employment and the frustration of attempting to live from day to day, which is now festering in the central cities, provide all the necessary ingredients for a tragic rerun of the crisis of the 60s as well as providing the impetus for increased crime and its concomitant social costs. Data from the National Longitudinal Study (NLS) of the High School Class of 1972 sponsored by the National Center for Education Statistics (CES), illuminates this point. The NLS is a long-term educational research program designed to track a national probability

[2]In 1975 the median income for black families rose to 61 percent of the median income for whites However, the gain was due to a smaller income gain for whites between 1974 and 1975 than for black households Median income for whites increased by only 6 percent in contrast to 10 percent for blacks (20) Moreover, white female income traditionally has been lower than income for white and nonwhite males, with black females reflecting the lowest wages (24)

sample of 23,000 young men and women who graduated from high school in 1972. The intent is to examine their educational and occupational patterns, plans, aspirations, and attitudes and to relate such information to the sample's prior educational experiences and their personal and socioeconomic characteristics. A recent report (13), noted that unemployed whites tended to say that they were either going to school or did not want to work; both blacks and those of Spanish backgrounds more often mentioned a shortage of jobs, inadequate training, or lack of experience as reasons for not working.

Enrollment statistics, like employment data, also provide a social indicator of the extent to which we are meeting the needs of our people. An examination of college enrollment data reveals that the enrollment of women and minorities has increased dramatically in recent years. Women now constitute about half of the first-year graduate enrollment in most institutions of higher education, according to 1975 US Census Bureau data (9). Between 1970–1975, the enrollment of women in graduate and professional schools rose about 75 percent, while the enrollment of their male counterparts increased only 23 percent. Primarily reflecting the impact of large enrollments, the number of women receiving doctorates increased 59 percent—from 4,600 in 1971 to 7,300 in 1975 (male doctorates declined 2.6 percent during this period, from 27,500 to 26,800); women's first professional degrees, in the same period, rose dramatically by 184 percent to a total of almost 7,000. Yet, statistics on declining enrollments after the first year suggest that women are less likely than men to receive their bachelor's and advanced degrees. Moreover, in terms of the number of doctorates conferred in selected fields, recent studies by the National Center for Educational Statistics show that in 1975, women received only 21 percent of the awarded doctorates, 13 percent of the first professional degrees in medicine, and 15 percent of the first professional degrees in law (14).

Enrollment data for minorities also showed a marked increase. An Office of Civil Rights survey showed that minority enrollment in colleges and universities rose 11.7 percent between 1972–1974, while total enrollment in the same period increased by only 2 percent. These increases were across the board for all disadvantaged minority groups—American Indian, black, Asian-American, and Hispanics. In 1972, the proportion of minority-to-

majority full-time total enrollment was 11 9 percent but by 1974 it stood at 13.1 percent. Freshman year statistics for 1974 showed that blacks constituted 10.8 percent of the total full-time enrollments; Spanish-surnamed, 3.5 percent, Asian-American, 1.1 percent; and American Indian, 0.7 percent. Moreover, from first-year undergraduate education through doctoral-level studies, black females consistently comprised a greater proportion of the total full-time enrollment than black males. More recent data suggest that minority enrollment has peaked and may be on the decline at predominantly black and white undergraduate campuses (19). Blacks have shown gains in enrollment in the law and medical schools: a 111 percent increase in enrollment in law schools in 1974 as compared with 1972 and an increase of 50 percent in medical schools.

In contrast, the degrees conferred have not kept pace with minority enrollment. For example, in the 1973–74 school year, blacks were estimated to have received only 5 1 percent of the bachelor's degrees awarded (8). Moreover, the percentages of ethnic minorities joining the professional ranks such as lawyers and physicians remain small and unfortunately stable. For example, 2.2 percent of all physicians were black in 1950 and in 1970 (2).

If one were to summarize the status of women and minorities in employment and college attendance, it could be safely concluded that impressive gains have been made in some aspects, but because of their low status to begin with, the gains have not been sufficient to allow for significant progress in representation in employment and higher education for these groups. And in some areas, particularly the job training of the older teenager or young adult, valuable ground has been lost

The Response

The executive and legislative branches of the federal government have responded to the needs of individuals and groups that have been the targets of discrimination through various laws designed to assure their protection. Laws against employment discrimination were written into the Civil Rights Act of 1866 and 1870 and the Equal Protection Clause of the Fourteenth Amendment. The Civil Rights Act of 1964 placed the federal government in an affirmative-action posture by providing the element of enforcement so sorely missed in

Brown I and II (28); under Title IV of this act, the U.S. Commissioner of Education is authorized to assist LEAs to desegregate and the Attorney General is empowered to institute lawsuits to bring about desegregation; Title VI forbids the use of federal funds in any federally financed program (such as apprenticeships, training, work-study) that practices racial discrimination, Title VII, as amended by the Equal Employment Opportunity Act of 1972, prohibits discrimination because of race, color, religion, sex, or national origin in any terms, conditions, or privilege of employment, including employment in educational institutions.

Other forms of antidiscrimination statutes include a) prohibitions against allocating differential pay on the basis of sex for the performance of similar work subject to the Fair Labor Standards Act (FLSA) as well as protection for other personnel in executive, administrative, and professional positions not covered by FLSA (the Equal Pay Act of 1963), b) stipulations that require the establishment of affirmative action programs by all federal contractors and subcontractors with contracts in excess of $50,000 and 50 or more employees, with the contractor subject to monitoring by an assigned federal compliance agency (Executive Order 11246 as amended by Executive Order 11375); and c) an extension of coverage of the Equal Pay Act that prohibits sex discrimination against employees or students of any educational institution receiving federal financial aid (Title 9, Education Amendments Act of 1972). Moreover, most states and many local government agencies have established laws prohibiting employment discrimination.

Moreover, state and local governmental agencies, colleges and universities, and some businesses began developing programs designed to increase minority representation.

Resolving Conflict: Recent Court Decisions

It was inevitable that the age-old conflict between justice and equality would finally reach the courts. Only a brief description of the *Bakke* or so-called reversed discrimination case should be required in light of its widespread publicity and gravity (3). A year after its opening in 1968, the University of California Medical School at Davis initiated the Task Force Program, a special admissions program that, between 1970 and 1974, admitted 71 minority

students: 26 blacks, 33 Chicanos, and 12 Asian-Americans. Forty-nine additional minority students were admitted through the regular admissions process during these same five years—one black, one American Indian, six Chicanos, and the remainder Asian-Americans (4). Sixteen percent of the available places in the entering classes were reserved for the special admittees, applicants judged to be disadvantaged by the special admissions committee on the basis of a number of personal characteristics among which, as conceded by the university, included race. While certain elements of the admissions process were unclear,[3] several aspects were apparent: At least for the 1973 and 1974 academic years, the undergraduate grade-point average (GPA) in science as well as the overall GPA for the special admittees were lower than for regular admittees, the average scores on the Medical College Admissions Test (MCAT) were consistently lower for the special program participants,[4] considerable overlap in college GPA (overall and science grades only) was evident between the two groups of admittees, no white student had ever been admitted (although it is unclear whether any ever applied) to the program since its inception, and apparently the applicants for the special program competed only among themselves for entry rather than in a common applicant pool. What is also apparent but not widely publicized is the fact that on the basis of the hard data—

[3]Indeed, the amicus brief filed by the Department of Justice pointed to several ambiguities (e.g., whether the final bench mark ratings of the special admittees were ever compared with those for the regular admittees and precisely how race was employed in the admissions process) in the university's admissions procedure in its argument for limiting the legal question before the court to the issue of whether the University of California Medical School may consider the race of applicants for the purpose of operating a properly administered affirmative action program (5).

[4]Scores for both groups of applicants were presented in percentiles, at least in the Department of Justice's brief (3), but assuming roughly a normal distribution and a standard deviation of approximately 100 on both subtests, it appears that the special admittees were roughly one standard deviation below the regular admittees on the verbal subtest for each year and for the years 1973 and 1974, approximately 1.33 standard deviations and 1.25 standard deviations, respectively, on the *MCAT* science subtest. These estimates of differences between the two groups, particularly for the verbal scores, are consistent with other studies of standardized aptitute test score differences between minorities and nonminorities.

that is, the traditional measures of college GPA and MCAT scores—Mr. Bakke *exceeded a large number* of *the regular admittees* as well. Therefore, the very question of status that prompted Mr. Bakke to seek redress in the courts appears to have been precipitated by the use of soft data such as the results of interviews and letters of recommendation. This final point may be moot with regard to the constitutional question of whether race can be used validly as one of the criteria for admissions. It is, however, terribly important to supporters of special programs who grow suspicious of the reasons why the program was singled out as the culprit when many factors in the rating process were equally vulnerable to questions concerning their relevance or validity in determining Mr. Bakke's final rating.

It is not the intention here to argue the merits of the case but to emphasize the fact that the residual effects that have been generated in terms of suspicions, hostility, and so on over the *choice of cases* to settle this critical issue may match those predicted this morning if preferential programs prevail.[5] However, I [illegible]es illustrate the extreme vulnerability of such programs to the kind of careful scrutiny they may expect even if the judgment of the Supreme Court of California is vacated.

The case before the Supreme Court was foreordained. The setting was established at least six years ago. The actors had only to be brought forward to center stage. In 1974, a similar case before the Court was rendered moot because the individual in question finally was admitted and was near graduation (12). However, even though this case was not decided, the principle of employing racially neutral admissions procedures was voiced by Justice Douglas' minority statement. Moreover, decisions similar to those handed down by the California court in the *Bakke* case were handed down by the New York Court of Appeals last year and the U. S. District Court, Southern District of New York, this year in cases involving special minority-student programs at medical colleges (1). More currently (and perhaps more ironic), the Colorado Supreme Court last spring took under consideration a white student's claim that the University of Colorado Law School illegally *excluded him from its special ad-*

[5]Indeed, the university's appeal to the U.S. Supreme Court over the objections of civil rights leaders, labor unions, and public advocates that the case was not well developed has raised some suspicion that a concerted effort was made to apply the coup de grace to special admissions programs

missions program because of his race (11). In this instance, the question was raised whether an educationally and culturally "disadvantaged" white student has a constitutional right to be considered for admission under the special program designed to give preference to minorities, mainly blacks, Chicanos, and American Indians. Finally, many of the current debates on the legal issues, social implications, problems, and benefits of preferential collegiate admissions programs were embodied in the papers published by the *University of Toledo Law Review* in 1970 (26).

More Conflicts and Possible Implications for Affirmative Action

The following two excerpts, the first from the dissenting opinion of Justice Harlan in *Plessy v. Ferguson* and the other from the affirming opinion of a federal district court in *Associated General Contractors of Massachusetts, Inc. v. Altshuler* (490 F. 2d9, 1973) concerning the "Boston Plan," an affirmative action plan which required contractors to employ a stated percentage of minorities, point up the issues that have arisen in recent years over the implementation of affirmative action programs:

> In respect of civil rights, common to all citizens, the Constitution of the United States does not, I think, permit any public authority to know the race of those entitled to be protected in the enjoyment of such rights . . . Our Constitution is color-blind and neither knows nor tolorates classes among citizens

> It is by now well understood, however, that our society cannot be completely color-blind in the short term if we are to have a color-blind society in the long term. After centuries of viewing through colored lenses, eyes do not quickly adjust when the lenses are removed. Discrimination has a way of perpetuating itself, albeit unintentionally, because the resulting inequalities make new opportunities less assessible. Preferential treatment is one partial prescription to remedy our society's most intransigent and deeply rooted inequalities

Our nation, wrestling with the consequences of its own past, has been and is still attempting to plot a course through the Scylla of the effects of racism and the Charybdis of individual injustice; of attempting to assure equality of opportunity and yet to preserve our

democratic form of meritocracy. In this struggle, there are many who express the view that the two opposing sources of conflict are more apparent than real; that our democratic tenets and aspirations dictate nothing less than the elimination of race as a factor in bringing about equality of opportunity; that preference on the basis of race is always discriminatory, probably invidious, and perhaps unconstitutional; that, indeed, "justice and equality are more or less incompatible" (7). Others noted that for 90 percent of the time that our great nation has existed, it has had institutionalized racism in some form or another and that the relatively short period since *Brown* cannot be viewed as sufficient for removing the effects of past racial and sex discrimination nor can it provide an equitable basis for minorities to compete with the majority now of in the near future. Race may be a proxy for other sins but it can also stand on its own merits. Racial discrimination, it is reasoned, perpetuates itself and requires more than laws and benign banalities. Others view the nation as tiring of the rough course and seeking refuge in retrenchment or, worse, retrogression.

Grounds for Agreement

As the opponents and proponents state their arguments, there are some grounds of apparent agreement. One is that *Bakke* could have implications far beyond the matter of college admissions, perhaps striking at the heart of affirmative action. Two: The decision, whatever it may be, will not absolve the university of its obligation to determine which course it should take in fulfilling its mission in a democratic society. This is as it should be, for it is the prerogative of these institutions to establish their educational policies. However, if the decision is in favor of Bakke, there is the increasing possibility of additional lawsuits. Since to many, discretion is the better part of valor, institutions of higher education may retreat entirely from special collegiate programs. Third: Sincere opponents and proponents agree that preference without regard for ability or personal qualifications is dysfunctional. However, there is lively debate over what is meant by "qualifications" or what the qualitative teaching role of the college should be. Opponents of affirmative action consider the affirmative action effort as that of a national discriminatory program that has had little effect on changing the status of racial minorities

and women because it rests on tenuous principles: Goals, no matter how they are explained by government bureaucrats, are viewed as quotas which, because of an undersupply of qualified individuals, can only be met by recruiting unqualified persons. Moreover, such programs will reap scorn and additional feelings of inferiority upon the individuals for whom the programs were designed to assist and cause greater class strife as more and more interest groups come forward demanding their fair share (15). One of the most frequently occurring responses from academia is that the affirmation process is abysmally low in cost efficiency, is time-consuming and unproductive in terms of paper work, and, worst of all, infringes on the autonomy of the university and debases the quality of the institution. Therefore, it is safe to assume that a decision in favor of Bakke would be considered as going a long way to redirect the country on a proper path toward true equality. Unfortunately, the programs or policies that have been offered—and they have been only a few—seem distressingly like afterthoughts and at best are totally unrealistic.

Supporters of affirmative action, and therefore for the overturn of *Bakke*, rely heavily on the compelling interest of the state in upgrading skills and increasing the representativeness among minorities and women in a wide spectrum of educational and business institutions. Such considerations as ratios of one physician for every 3,500 black citizens as opposed to a 1,750 ratio for whites, or one black attorney for every 6,000 black persons in contrast to a 1,630 ratio for whites, prompt supporters to advance the notion that it is in the state's interest to support aggressive affirmative action programs. Benign or "color-blind" methods are viewed as totally unrealistic, there are no adiaphorous remedies. Given the fact that evidence of beneficial impact is now appearing, supporters view the idea of benign neglect as tragic. Moreover, supporters are vociferous in their distinctions between goals and quotas. The former are ameliorative and the latter repugnant, goals are only goals—targets to aim at through good-faith efforts with more consideration given to qualifications or possible remediation than to a mere numerical count. Since bright critics are having difficulty making the distinction between goals and quotas, it is felt that there is little sincerity, only the opportunity to use the term as a cudgel to strike down the total affirmative action effort. Moreover, in terms of education, it is felt that the university should be a microcosm of the society of which it

is a part. Therefore, the representation of minorities and women in universities is in the interest of the state, the university, and its student body. In regard to discrimination, reverse discrimination is thought to be benign, temporary, and possessing one important characteristic that distinguishes it from traditional discrimination: It is employed by the majority to *include* rather than *exclude* minorities in the realm of human activities that have been enjoyed by the majority, particularly white males (16). Obviously, for affirmative action supporters, a ruling for Bakke would be viewed as having disastrous consequences.

In this author's view, a decision for Bakke would be a serious blow to the progress we have made thus far as well as a reinforcement of the growing belief that a national backlash is now taking place. Unfortunately, the greatest impact of such a decision may be on those 15 or so federal domestic programs designed to give assistance to those who need it most. For example, the $4 billion public works law passed by Congress last spring in an attempt to stimulate employment in the construction industry requires that 10 percent of the money for each construction project be spent on purchases from minority businesses. Just a month ago, the Administration ordered government agencies to double purchases from minority-owned businesses during the next two fiscal years (27). Moreover, several states and cities have been required by the federal courts to add more racial minorities to their payrolls with hiring ratios being a part of the order. It would seem that all of these efforts would be in serious jeopardy.

Conclusion

In this paper, I have discussed only a few of the problems facing us in our efforts to advance equality of opportunity. I realize also that I have not mentioned problems within two areas of my own research interests—the effectiveness of collegiate compensatory programs (there are many, and they no longer can rest on the beneficent character of their mission but must be held accountable for their efforts) and the problems of admissions tests (I am deeply disturbed over what appears to be an ever-increasing level of importance given to small score differences on graduate school admissions tests where the competition is fierce and the decision is often relegated to a mere quantitative "pecking order."). However, it would seem that among

the many aspects related to the social implications of the pending court decision, it will be its broad impact upon society that will be of major interest to future historians. Future observers may conclude that the height of this nation's greatness may be measured by the depth to which we refuse to allow our most disadvantaged citizens to descend.

References

1. *Alevy v. Downstate Medical Center*, New York State (1976). See also *Chronicle of Higher Education*, July 5, 1977, 7, for report of the U.S. District Court for the Southern District of New York ruling.
2. Bell, G. Brief Amicus Curiae for the United States in the Bakke case, 10:45.
3. See, for example, Bell, *id.*, for statement of the facts as presented by the federal government or the *Bakke Decision—Disadvantaged Graduate Students*, Committee Report (draft), Assembly Permanent Subcommittee on Postsecondary Education, California Legislature, July 1977, for presentation by the State of California
4. Bell, *id.*, at 9.
5. Bell, *id.*, 68–74
6. Brimmer, A. F. Income distribution and economic equity in the United States. Paper read at the 1976 annual meeting of the American Association for the Advancement of Science, February 1976
7. Brubacher, J. S. *On the philosophy of higher education* San Francisco Jossey-Bass, 1977, 56–67
8. Bush, S. *Minority group participation in graduate education*. Washington, D.C.: National Board on Graduate Education, 1975.
9. *Chronicle of Higher Education*, February 7, 1977, 13 No. 21, 1, 10.
10. *Chronicle of Higher Education* November 8, 1976, 13, No. 10, 7
11. *Chronicle of Higher Education*, 14, No. 2, March 21, 1977, 6
12. *De Funis v. Odegaard*, 414 U.S. 312, 340 (1974)
13. Grant, W. V., & Lind, C. G. *Digest of education statistics, 1976 edition* Washington, D.C.: National Center for Education Statistics, 1977, 182.
14. Grant, *id.*, at 123 Table 114

15. *See* B. R. Gross (Ed.) *Reverse discrimination*. Buffalo, N.Y.: Prometheus Books, 1977; A. Etzioni. Making up for past injustices: How Bakke could backfire. *Psychology Today*, August 1977, 18; and a special issue of the *Urban League Review*, Summer 1977, 2, No. 2, for a divergence of views on affirmative action.

16. Gross, *id.*

17. Hayghe H. Marital and family characteristics of the labor force in March 1973. *Monthly Labor Review*, April 1973, 24–25

18. *Id.*, at 95, Table 97

19. Majer, K. *Minority underrepresentation in national science graduate programs*. La Jolla, Calif., University of California, San Diego, November 1975; see also the *Chronicle of Higher Education*: March 7, 1977, 1, for report of enrollment decline among the 105 predominantly black colleges.

20. National Urban League. The economic status of blacks—1975–76. *The Urban League Review*, Summer 1977, 2, No. 2, 52–57

21. U.S. Commission on Civil Rights. *Last hired first fired: Layoffs and civil rights*. Washington, D.C.: U.S. Commission on Civil Rights, February 1977.

22. U.S. Commission on Civil Rights, *id.* 8–9

23. U.S. Commission on Civil Rights, *id.* 8–9

24. U.S. Commission on Civil Rights, *id.* 81

25. U.S. Department of Commerce, Bureau of the Census. *Social and economic status of the black population in the United States*. Current populations report. Washington, D.C.: Department of Commerce, 1974.

26. University of Toledo Symposium. Disadvantaged students and legal education—programs for affirmative action. *Toledo Law Review*, 1970, Nos. 2 & 3, 1970

27. *Wall Street Journal*, September 13, 1977, 7

28. Wilberg, W. *Minority students: A research appraisal*. Washington, D.C.: U.S. Department of Health, Education and Welfare, National Institute of Education, 1977

There Ought to Be a Law

Norman Frederiksen
Senior Research Psychologist
Educational Testing Service

A good many people are apparently very much concerned these days about tests and testing practices, some people even think there ought to be a law. Bills are being introduced in both state and federal legislatures to control testing practices. Congressman Harrington, for example, is sponsoring a bill that, among other things, would require putting a warning label on all score reports stating that "These scores are approximations" and including a report of the standard error of measurement.

The National Education Association has recently reaffirmed its opposition to standardized tests, and the report of the NEA task force provides some reasons for its stand. Tests are thought to be deficient because they measure cognitive learning to the exclusion of emotional and physical development, they penalize creative thinking because of their heavy emphasis on multiple-choice items, and they are often culturally biased. There are other frequently voiced criticisms that NEA might have included, such as that tests are coachable, that they influence the curriculum, and that they predict only narrow academic criteria and not career success.

Some people in the testing business tend to reject such criticisms out of hand by saying that the critics don't really understand the problems, or that any faults of testing are attributable to misuse of tests rather than to deficiencies in the tests themselves. But I think it is a good time for those of us involved in educational and psychological testing to stand back and take a fresh look at what we are doing. There must be reasons other than political expediency for the widespread protest. Maybe we ought to consider more seriously what improvements in tests and testing procedures should be made before the law imposes restrictions that make it more difficult to move ahead in the testing field.

Some of the criticism can be disposed of quite easily. *Of course* a

cognitive test doesn't measure emotional development. Neither does a thermometer measure humidity, but the thermometer is still useful.

Other criticisms are not so easy to dismiss. For example, the idea of reporting the error of measurement along with test scores has been around for a long time, and some tests do use score report forms that call attention to the error of measurement. But the great majority of score reports are not accompanied by any kind of graphic, numerical, or verbal indication of the error in measurement. True, the test manual may contain a section on error of measurement, but it is easily overlooked or forgotten when score reports are examined. Why isn't it more common to put the information on the score report itself? Do we have to be forced by law to do something that most test experts would consider sound practice?

In the short time I have this afternoon, I can't go very deeply into a discussion of which criticisms are justified and which ones are not. But I would like to comment on the following allegations about tests: that multiple-choice tests penalize creative thinking; that tests don't predict success beyond narrow academic criteria; and that tests influence the curriculum.

Do Tests Influence the Curriculum?

Let's start with the charge that tests influence the curriculum. A number of years ago, a superintendent of public instruction in a state not far from here stated publicly that he favored banning all educational tests because they influence the behavior of teachers. I thought he was wrong in wanting to ban tests. If I had been the superintendent and I had in my grasp a tool that I could use to influence the behavior of teachers, I would have held on for dear life. Let me tell you why I think he missed the boat.

During World War II, I was a staff member of a project, headed by Harold Gulliksen, that was supposed to do research on selection and training of naval personnel. Among other things, we conducted validity studies of the tests used in assigning recruits to naval training schools. One of our findings was that the best tests for predicting grades in gunner's mate schools were verbal and reading comprehension tests. This didn't make much sense, in view of what gunners mates are supposed to do, but assignments to service

schools were nevertheless made in accordance with the empirical evidence.

Later, our research group was given the assignment of improving grading practices in the navy service schools, and I was sent to work at the gunner's mate school in Bainbridge, Maryland. We found that the lecture-demonstration method of teaching was used. The students studied the technical manuals, and the examinations consisted of multiple-choice items based on the lectures and manuals. The items dealt with such topics as muzzle velocity and the function of the breech block locking bolt. Since the job for which these students were being trained was to maintain, adjust, and repair the guns aboard a warship, it seemed more reasonable to use performance tests. Accordingly, we developed a set of tests that required students to perform such tasks as adjusting the oil buffer for maximum rate of fire on a caliber .50 Browning machine gun, removing and replacing the interlock carrier spring from a 20 MM gun, and removing and replacing the extractor plunger on a 5"/38 antiaircraft gun. The instructors complained that the tests were too hard. They were right. Few of the students could perform the tasks, even with liberal time allowances.

Since we had orders, the performance tests were nevertheless given at the end of each unit of training. Because successive classes overlapped, the new students soon got word as to what the new tests were like, and they began practicing the assembly and disassembly of guns. Performance on the tests improved with each class. The instructors also got the point. They moved out the classroom chairs and the lecture podium and brought in more guns and gun mounts. The upshot was that students spent most of their time practicing the skills required in repairing and adjusting guns. The tests soon became too easy. The validity coefficients changed too: The verbal and reading test validities dropped, and the mechanical aptitude and mechanical knowledge tests became the best tests for predicting grades in gunner's mate school (4).

Note that no attempt was made to change the curriculum or teacher behavior. The dramatic changes in achievement came about solely through a change in the tests. The moral is clear. It *is* possible to influence teaching and learning by changing the tests of achievement. It is also clear that those who make the tests have a great responsibility to produce tests that influence teachers to teach, and students to learn, the knowledge and skills that truly reflect the ob-

jectives of the training program.

While I am on the topic, I should mention the related issue of coachability of tests. When important administrative decisions depend on test scores, students are likely to seek to gain an advantage by attending coaching schools or buying books on how to take tests. If teachers think *they* are to be evaluated on the basis of their students' scores, they are tempted to give sample tests for the students to practice on. Such coaching is usually viewed as undesirable, and no doubt it is undesirable under some conditions. But note that what went on in the gunner's mate school was basically coaching for the tests. Coaching is undesirable only if it results in improving test scores *without* increasing proficiency with regard to the real instructional objectives. If tests can be made that truly reflect the school objectives, the difference between coaching and teaching disappears, and efforts to get high test scores are the same as efforts to attain the desired skills and information. We should be making achievement tests that are coachable in this desirable sense.

To summarize: Yes, tests do influence the behavior of teachers and students. The only question concerns what the teachers and students are influenced to do. We should be providing tests that encourage the teaching and learning of the knowledge, skills, and abilities that represent all of the educational objectives, not merely those that are easy to measure with a paper-and-pencil test.

Do Tests Penalize Creative Thinking?

Now let's turn to the criticism that multiple-choice tests penalize creative thinking. Presumably this charge is based on the fact that in taking a multiple-choice test, the examinee does not have to think of the options for himself, he merely has to read the options, evaluate them, and choose the best one. These activities require comprehension of the problem, a background of relevant information, and a certain amount of reasoning and judgment, depending on the nature of the particular item. Real-life problems more often present themselves in an open-ended form. A problem arises, and the question is "What shall I do?" The individual must then think of at least one solution to the problem. Better, he will think of several, or many, possible answers. From here on, the process may be much the same as for multiple-choice problems. Reasoning and judgment

are involved in choosing the final solution. The necessity of thinking of the options for oneself appears to be what characterizes a free-answer problem, and thinking of the options presumably requires such abilities as originality and ideational fluency—the divergent-production abilities in Guilford's structure-of-intellect theory (2).

Bill Ward and I have been gathering data to find out what psychological processes are involved in problem solving, particularly creative problem solving (1). We have developed a set of scientific thinking tests intended to provide the dependent variables for studies of problem-solving behavior. These tests are meant to simulate some of the tasks often performed by a behavioral scientist. Their titles are Formulating Hypotheses, Evaluating Proposals, Solving Methodological Problems, and Measuring Constructs.

I'll describe Formulating Hypotheses. Each item is a brief description of a scientific investigation. The results of the investigation are shown in the form of a graph or table, and the major finding is clearly stated. The examinee's task is to suggest hypotheses that might explain the finding. He is asked to write not only the hypothesis he thinks is most likely to be correct, but also other hypotheses that should be considered in interpreting the data or in planning another investigation. The test is not of the multiple-choice form. The candidate must think of the hypotheses, write them down, and then mark the one he considers most likely to be corr . In other words, the examinee has to think of the options for self before choosing the best one.

A number of scores can be obtained from a test composed of such items. We have worked so far with six scores: (1) the average quality of all the hypotheses written by a candidate, (2) the average quality of the hypotheses the candidate thinks are his best, (3) the average quality of the hypotheses that are best according to our scoring system, (4) the number of hypotheses written, (5) the number of unusual hypotheses, and (6) the number of hypotheses that are not only unusual but of high quality. Our investigation of the psychometric properties of the tests shows that reliable scores can be obtained, that the quality scores are not highly correlated with number scores, and that the tests are of suitable difficulty for candidates for admission to graduate school or for first-year graduate students.

We have also developed another form of the Formulating Hypotheses test that is scorable by a machine. In this version, a list

of hypotheses is presented, and the candidate is given two tasks: First, he is asked to mark those hypotheses on the list that he thinks ought to be considered and, second, he is asked to choose the best hypothesis on the list. Scores analogous to those used for the free-response form are obtained.

We are now analyzing our data to find out how the free-response and machine-scorable formats differ with regard to what they measure. One finding is that the correlations between corresponding scores from the two forms are low. The highest correlation is .33; this is the correlation between forms for scores based on the quality of the ideas the candidate thinks are his best. The correlation between forms for number of hypotheses is .19 and for number of unusual hypotheses it is .17. These low correlations are not attributable to low reliability. Clearly, the free-response and machine-scorable tests do not measure the same thing.

Another kind of comparison has to do with relationships of the scores to measures of various cognitive abilities. The results show that the *quality* scores are related to the same kinds of ability, whether the scores are based on free-response or machine-scorable forms: Quality scores from both forms correlate with tests of reasoning abilities, particularly inductive reasoning and logical reasoning, and also with tests that measure cognitive flexibility—that is, ability to change sets in solving a problem. But the two forms differ strikingly with respect to the correlations of number scores with measures of divergent production. Only for the free-response form do the *number* and *number of unusual* responses correlate significantly with divergent-production measures—tests of expressional fluency, ideational fluency, and originality.

These findings are based on some preliminary results just off the computer. A more comprehensive analysis will be made and reported later. But it appears that the answer is clear, at least for the two forms of the Formulating Hypotheses test. If you are interested only in the quality of the ideas, the free-response and the machine-scorable test do not seem to differ appreciably with respect to the cognitive abilities required. But if you are interested in the candidate's ability to think of options for himself, and to think of options that most other candidates do not think of, differences are found—the free-response test taps divergent-production skills that are not measured by the machine-scorable form. To the extent that tests influence the behavior of teachers and learners, free-response tests

would presumably be more likely to enhance the learning of the divergent-thinking skills that are involved in problem solving.

Do Tests Predict Career Success?

Now let's deal with the criticism that tests may predict grades but they do not predict career success. Good data on this topic are hard to find, primarily because satisfactory criteria of career success are not usually available. Most studies of career success use such criteria as supervisors' ratings, salary, or rate of promotion, all of which are as likely to reflect how well liked or popular the individuals are as how proficient they are in whatever kinds of problem solving and decision making may be required of them. If we want to know how well tests predict career success, we need better criteria of success. And this probably means that we must create them.

The Formulating Hypotheses test and the other scientific thinking tests were originally developed to provide criterion variables relevant to the work of a behavioral scientist; so perhaps we should see how well conventional selection tests predict scores on these tests. We find that the correlations of GRE aptitude and achievement tests with scores on the free-response tests of scientific thinking are not high. The median correlations of quality scores with the three GRE tests (V, Q, and Advanced Psychology) are approximately .35. For the scores based on number of responses, the median correlations with GRE scores are .20 or lower. Assuming that the tests of scientific thinking are reasonable simulations of aspects of the work of a behavioral scientist, clearly the conventional tests predict only those aspects of scientific thinking that involve reasoning, not those that require the divergent-thinking abilities presumably involved in creative problem solving.

Some findings from research on medical education are relevant to the issue of predicting career success. There is a good deal of evidence that conventional aptitude and achievement tests are reasonably good predictors of grades in the first two years of medical school, when students are getting heavy doses of anatomy, histology, biochemistry, and so on. But the same tests are poor predictors of success in the clinical years, when students are learning to deal with patients, make differential diagnoses, and develop plans for patient management. True, the evaluations of clinical ability are

based on ratings and on student performance rather than career performance. Nevertheless, the failure of conventional tests to predict success in clinical training at least suggests that they would also fail to predict clinical performance in medical practice.

Because of the concern of medical school people that the conventional selection methods may not predict clinical skills, a research study was initiated that involves development of criterion measures. The schools were particularly interested in the problem-solving skills required in medical diagnosis and interpersonal skills of the sort required in interviewing a patient. The study is being carried out cooperatively by a consortium of medical schools, ETS, and the National Board of Medical Examiners.

The National Board of Medical Examiners has primary responsibility for developing the criterion measures. The diagnostic problem-solving tests they have developed are paper-and-pencil simulations of what happens when a doctor meets a new patient, hears the patient's complaint, interviews and examines the patient, orders laboratory tests, and so on, until a differential diagnosis is made. The examinee has opportunities to indicate the diagnostic hypotheses that come to mind, the information he needs to evaluate the hypotheses, the hypotheses he entertains after more information is provided, and so on, until the decision is reached. The tests are appropriate for fourth-year students and residents. Both free-response and machine-scorable versions of the tests have been developed.

The criterion tests in the area of interpersonal skills require that the physician interview live, simulated patients. Each "patient" has been trained to describe his or her complaint, to reveal certain other information if asked, and to show appropriate affect. Scores will be developed after coding the many behaviors occurring during the tape-recorded interview. The scores will reflect the physician's performance in a great variety of ways, such as responding to the patient's affect, time spent asking questions, time spent giving information, time spent listening, number of interruptions, and so on.

Given such a set of criterion measures, it seems reasonable that the most valid selection tests would be counterparts of the criterion tests that are suitable for candidates who have no medical training. Educational Testing Service has developed a set of experimental selection tests of this kind. The problem-solving tests involve social, education, and ecological problems, and they similarly provide op-

portunities for the examinee to indicate what hypotheses he is considering at each of several stages in dealing with a problem, what information he needs, what sources of information he may try to exploit, and so on, until a final solution is called for. Again, both free-response and machine-scorable formats will be employed.

For the interpersonal skills tests, the counterparts of the simulated patients will be simulated counselees with various kinds of personal problems that a candidate for admission to medical school could deal with. Analogous methods of scoring will be used. The tests will be administered to first-year and fourth-year medical students.

What a Test Is

As you may have noticed, I have been trying to suggest by my illustrations that we have too narrow a conception of what a test is. To most people, a test is a booklet with a lot of multiple-choice items and a separate answer sheet. Let me propose a much more general definition: A test is any standardized procedure for eliciting the kind of behavior we want to observe and measure. I mean the behavior we *really* want to measure, not merely something related to it. This definition imposes few constraints on test makers. If we want to measure spelling ability, we can dictate words to spell. If we want to measure ability to repair machine guns, we can present guns needing repair. If we want to measure diagnostic ability, we can provide opportunities to deal with patients who have problems. If we want to measure interviewing skills, we can present someone to be interviewed. Such tests would presumably influence teachers and learners, and be coachable, in the desirable sense: they would require the candidate to think of the options when this is appropriate; and they would have a better chance of predicting career success than conventional tests.

I'm sure you are all thinking that that is all very well, but how can you test a million candidates a year with tests that may require machine guns or simulated patients, tests that must be administered individually and that can't be scored with an optical scanning device?

One answer is that perhaps it is time to recognize that assessment and evaluation may be sufficiently important, at least in some

instances, to justify much more than a day of testing time and a fee of fifteen dollars. If, for example, we consider the cost of a medical education, the additional income over the lifetime of a practicing physician that is attributable to his medical training, to say nothing of the importance of the life-or-death decisions made by a physician, the cost of conventional testing is small indeed.

Another answer is related to what I have said about coachability and the influence of tests on teachers and students. I have argued that it is possible to make tests that reflect instructional objectives more accurately than do conventional tests and that such tests influence the behavior of teachers and students in ways that enhance learning. If I am correct, it would seem sensible to use tests for teaching, not just for evaluation. Forms of a test could be constructed in such numbers and variety that they could be used regularly for homework or classroom drill. Students could cram and teachers could coach as much as they pleased. The cost of the tests would be justified by their value for instructional purposes.

The scores would be useful in other ways, provided that on some occasions the tests were administered under standard conditions and records of performance were preserved. The scores could be used in monitoring the progress of individual students, for school evaluation, for counseling, for admission to more advanced training programs, and for assigning grades to students. School grades are now the best predictors of academic success in higher education; school grades based on tests that more adequately reflect the spectrum of educational objectives might also become the best predictors of career success.

I realize that I have oversimplified the issues and ignored some thorny problems and that my proposals are vulnerable to attack from several practical and methodological points of view. But I have been trying to make a point. I think we have drifted into stereotyped attitudes about tests and testing that have tended to freeze tests into formats that seriously limit their potential, and I want to encourage you to think of testing in the much more general sense that I have described. I think the consequence would be better tests—tests that exert a good influence on teaching, that are less susceptible to misuse and misinterpretation, that are viewed as "fair," and that are less likely to make people grumble that "there ought to be a law."

I discovered recently that Professor A. Lawrence Lowell, who was president of Harvard when I was born, had somewhat similar

thoughts about examinations. Here is a brief excerpt from an article he published in the *Atlantic Monthly* in 1926 (3):

> The question of studying for marks rather than for knowledge, and the kindred matter of cramming for examinations, are not uninteresting and are often misunderstood. The popular impression . . . is that a student whose primary object is a high grade devotes himself . . . to memorizing small, and comparatively unimportant, points in a course, and thereby makes a better showing than a classmate with . . . a larger real command of the subject. . . . As the [examination] questions are often made out and marked this result may, and does, occur. But if all examinations were so conducted as to be an accurate and complete measure of the education the course is intended to give . . . . then there would be no reason why the student should not work for marks, and good reason why he should. To chide a tennis player for training himself with a view to winning a match, instead of acquiring skill in the game, would be absurd, because the two things are the same . . . if marks are not an adequate measure of what the course is intended to impart, then the examination is defective. If examinations were perfect the results would command universal respect, and high grades would be a more general object of ambition.

References

1. Frederiksen, N., & Ward, W. C. Measures for the study of creativity in scientific problem solving. *Applied Psychological Measurement*. (In press)
2. Guilford, J. P. *The nature of human intelligence*. New York: McGraw-Hill, 1967.
3. Lowell, A. L. The art of examination. *Atlantic Monthly*, January 1926. Reprinted in *The Work of the College Entrance Examination Board 1901–1925*. Boston: Ginn and Company, 1926. Pp. 31–43.
4. Stuit, D. B. (Ed.) *Personnel research and test development in the Bureau of Naval Personnel*. Princeton, N.J., Princeton University Press, 1947.

The Logic of Judgment in Evaluation and the Law: Making Hard Decisions with Soft Data

MICHAEL SCRIVEN
Director, Evaluation Institute
University of San Francisco
Professor
University of California at Berkeley

Introduction

The way to make hard decisions with soft data is not, in general, to harden the data, although that may help: It is to improve the decision-making process, the methodology. And here I think we can learn something from the law. The last decade or two have seen the emergence of new models of legal reasoning, in particular the so-called New Rhetoric approach. In the same period, the discipline of evaluation has also taken on a new form or forms—it has become independent of testing and measurement and it has spawned its own new models, the Stake reactive model, CIPP model, Eisner's connoisseurship model, and so on. Evaluation is a new discipline that has emerged from testing in somewhat the way that new law emerges from sociology and demography. What I propose to do in these few notes is sketch in some arguments for supposing first, that this emergence is truly a revolution and second, that there is a similar logical structure underlying both legal and evaluative reasoning—at least, up to a point—a structure that goes far beyond the logic of quantitative inference that guides the measurement field, and that the essence of this structure is quite well expressed in the phrase "making hard decisions with soft data." We should, however, amplify the formula slightly, for hard decisions are just as hard, even if the data are hard if the *connection* (the relevance of the data to the decisions) is *soft*, as it often is.

The Way Things Were

The scenario that has unfolded on the methodological stage in these last years has introduced a radically different cast of characters. In 1960, the social sciences appeared on stage like a chorus of IBM junior management types—almost uniformly dressed in neat but boring replicas of the costume favored by their elders and betters, the Physical Sciences. That fellow down the end of the line there does look a *little* out of place—collar askew, shoes muddy and not tied—that's Anthropology, actually *Cultural* Anthropology. His older brother, Physical Anthropology, blends in perfectly, however. Nothing basically wrong with the family. If you look closely, too, you might wonder about the chap nearest you; he's primped up so exquisitely as to raise some doubts about his reality—that's Economics. Bang in the middle of the chorus, we can see a fine young fellow with a familiar look—why, it's Educational Tests and Measurements! What really impresses one about the whole bunch is their *spirit*, though. Just bubbling over with pride and enthusiasm. Reminds one a bit of that *Up With People*, show, don't you think? Another young fellow over there catches one's eye as an up-and-comer; what's that label say? Scientific Jurisprudence, by George! Definitely gives one a feeling that all's right with the world—and going to be better—with this splendid team ready to pick up the burden. Of course, there aren't any *girls* in the show, but you can certainly visualize them doing tremendously valuable things off-stage—sewing the costumes, cleaning up, and whatnot.

Well, it *was* a jolly good show. What was it called? Let's see—the Newtonian Follies of the Nineteen-Fifties. Still running in the Sixties, too, but they never changed the name—can't get the same alliteration, you know. And besides, the show itself never changed, so why bother?

Some of those young chaps were *definitely* born for greatness; like that Economics fellow. Went on to become a real star; they set up a special Nobel Prize for him, I believe.

That was the show, folks, and now it's folded, though not all the characters have quite realized it. What really happened was that people couldn't quite believe it any more, and if people don't believe in fairies, why, they just *die*. Its truly sad. But then, life must go on, and we have to grow up and give up our fairy tales.

The Methodology Theatre today has a really different show. No chorus line. Some women and minorities and even some oldsters on

the stage. Even the Anglos aren't dressed like each other. There are family groupings, but they're not easy to spot, and seem to be changing all the time. Truly bizarre characters have emerged, like that weird entity they call Policy Analysis, and the feisty newcomer Sociobiology. The Director says it's "experimental theatre," whatever that means. He says that playwrights today feel they have to start all over, that the old tradition isn't just dated, it's false. He says that the triumph of Economics now looks like the triumph of a bunco artist, with the Swedish Academy in the role of suckers. The magic of math models bemused them; the internal precision of the inferences kept them from noticing that the predictions were no good, the explanations were so facile as to be fraudulent, the conceptual schemes so frail or so fuzzy as to be fakery.

Now Tests and Measurements (T&M) was never a leader in that cast, but it surely was misled and it did its share of misleading when the show was riding high. It's worth examining its situation a little more carefully.

I don't suppose anyone would argue strongly for Tests and Measurements as any more than a set of tools, of means to ends. But, like their ally Statistics, they soon generated their own energy, their professional associations, their entrepreneurs and publications, and even their lobbyists in one or another (and finally in many) centers of power. The comparison with paleography with its numerical taxonomy in which the quantifiers got out of hand is illuminating, as is the analogy with mathematical economics. The substance of science was too quickly identified with the presence of quantitative variables and (more or less) operational definitions of them in terms of measurement procedures. Well, there's nothing new about the suggestion that too many crimes were committed in the name of norm-referenced testing, or analysis of variance, or market models. Let's try to be more constructive. There are two or three disciplines that point to an alternative approach, a more promising perspective for redemption. I have in mind *accounting*, *jurisprudence*, and *evaluation*.

The Way Things Should and Shouldn't Be

If you look at what has happened to the General Accounting Office, or HEW's Audit Agency, or California's office of the Legislative

Analyst, what you'll see is the massive dequantification of some parts of the job of some members of the accounting profession. It is a far cry from GAO's original mission to the investigation of the Tonkin Gulf incident—from auditing the books to the assessment of foreign policy decisions. Yet GAO has acquitted itself well at both ends of this long spectrum, as has Audit Agency. By and large, they have made the transition more gracefully than most of the quantitative generation in the social sciences, which still clings to an inappropriate model. Why? Perhaps it's because they see the practical necessity for change—the need of the decision makers for relevant research—and they are less diverted by the will-o'-the-wisps of a Newtonian synthesis, a great world model turning on the bearings of mathematics. Or perhaps it's because the quantities with which the accountants were preoccupied always had a highly pragmatic referent—money—while the social scientist has never quite been able to swallow the idea that true costs have any right to be regarded as legitimate variables. Least of all the bulk of economists, whose capacity for cost analysis is on a par with their interest in it. (The exceptions like Hank Levin stand out in such solitary splendor as instantly to validate the general truth.)

In any case, accountancy has transcended bookkeeping and has achieved a modest capacity for making hard decisions from soft data—that is, from all the data *other* than bookkeeping entries that must be sifted in order to achieve a reality-related recommendation. Nowhere is this better illustrated than in Abraham Biloff's brilliant book *Unaccountable Accounting*.

The situation in the law is very similar to that in accounting. The effort to quantify the *whole* decision making process has been more or less abandoned, since it can be done only by giving soft data the appearance of hard data and by making judgmental decisions into mathematical deductions, which instantly generates bad decisions. The subtleties of statistics, which can help us express *some* of the softness, cannot save us from the eventual need for holistic evaluative judgment in the synthesizing step where we have to trade off such disparate considerations as the chance of recidivism against the possibility of a mistaken verdict.

This is not to deny the colossal power of simple math models of decision making in the Meehl-Dawes tradition—models that make a shambles out of most claims for the ineradicable necessity of human judgment throughout the present territory. Most standard clinical

and legal judgments *could* be made better (especially given that equity is a value in itself) by a formula. But not *all* of them and especially not all of the really important ones. For those we need judgment, but *trained* judgment, not capricious judgment; and we should face the fact that the task of the social sciences in education is to improve that training, uncover and defuse the biases that operate, routinize what can be routinized, and assess the effectiveness, reliability, and validity of what's left.

Training, Not Just Testing

This move towards training and not just testing, towards improving the human instruments and not replacing it, can be seen in the spectrum of activities at ETS in recent years, and it can be seen in the new pattern of Supreme Court decisions since the desegregation decision. For all the effort towards strict constructionism, towards conservatizing the court, it still reaches farther into the social realm than ever before and in doing so, at once demonstrates the difficulty of the task and the recognition of its necessity. What we might call *social* jurisprudence has been edging out the "scientific" approach. And the pressure that has led to this is not political appointment—it is the wider perception of reality and its connections that the latter half of the 20th century has brought us. The law has absorbed the new precision of the quantitative disciplines well enough, but it has not been overrun by them. In the mainstreaming decisions, in the busing area, in *Serrano-Priest*, in *Bakke*, there is no lack of tests and measurement and statistics. But there remains a more important—though not very well conceptualized—model of law and the system of justice as a social change agent, a discipline that is not divorced from current events but is creating them and also studying the way in which it is creating them, self-referent jurisprudence.

This last feature is one of the most important distinctions between traditional conceptions of Tests and Measurements and the discipline of evaluation under which T&M must now be seen as largely subsumed. This very conference shows how far testing has followed the judicial system in becoming a change agent and becoming aware of it. But what part of the *training* in educational psychology has been shifted to the study of the social impact of testing? As yet, not much. Like testing, evaluations often have effects that are larger

than those of the program being evaluated: not hard, because few programs work, but not easy, because few evaluations are effective. Now T&M has long been aware of the problem of "reactive" or obtrusive measurement. Evaluation is extremely aware of reactive *reporting* as well and has long been studying and improving the report. It was also common enough in the social sciences to notice self-fulfilling or self-refuting prophecies, a kind of self-reference; but it was *not* common to take seriously the idea that it was an *obligation* of the investigator to investigate the investigation. The environmental impact research of the field biologist must also be studied by that same biologist for *its* effect; this is the intrusiveness imperative, the limiting case of *exhaustive* self-application of the scientific method. Evaluation has this very strong and distinctive emphasis on the *evaluation of evaluations*. This has also led to a sterner and wider view of reactive measurement, not only with respect to protecting privacy but in simpler ways: The very idea of a four-page instrument for the student's evaluation of teaching is now seen as absurd because of what it does to response rate and response stereotyping.

There are many other respects in which contemporary evaluation has grown a long way beyond its quantitative ancestors. It has developed new dimensions of analysis, never to be found in the T&M training programs. (Cost analysis is one of these.) Again, while it has been avidly soaking up alternative approaches to causal investigations, it has also totally transcended the notion that evaluation is either necessarily, typically, or principally a matter of causal investigation at all. The evaluation of curriculum materials for sexist bias or factual error via content-analysis techniques is not a causal investigation. The matching for congruence of test items with course objectives, the analysis of the sequencing of curriculum modules in terms of logical progression—these are noncausal evaluations of tests or curricula (or very large parts of such evaluations in some evaluation contexts). The creation of correct descriptions of complex programs may be the main task of an evaluation; it is not causal analysis. Looking for injustice in a school's admission or disciplinary system is not causal investigation. In even more general terms, evaluation has come to see that traditional control-group methodology is a limited special case even in causal research, not because there are better approaches involving no comparisons—there aren't—but because the *no-treatment* control is usually the

wrong one to use and always the wrong one to use alone. Instead, we have to *devise* or *uncover* the appropriate *comparison* groups, for decisions are rarely between treatment and no treatment but between two alternative treatments, and one or more of these may have to be invented by the evaluator if the evaluation is to be maximally useful to the decision maker. Thus, as with the law, creativity as well as study comes into full play.

Again, there is the evaluation of tests themselves, the activity which truly establishes the dominance of evaluation over T&M. Test people are becoming increasingly aware of and skilled in this, as you all know. What they are doing is evaluation, something much more general and less quantitative than most measurement. Finally, one might mention the extent to which evaluation has turned away from accepting the framework of intention, of goals, towards looking at the actual effects of a program or product.

The Need to Reach beyond Measurement

Rarely has the need for better tests been more obvious than in this meeting where we have reluctantly but seriously considered selection for admission to professional schools by *randomization* in the absence of better methods. It is clear that the evaluation of tests using only the standards of the discipline of tests leads to some of the worst abuses of testing—the confusion of correlation with common reference, of aptitude with ability, of criterion-referencing with competency, of objectivity with multiple choice, or reliability with consistency, of internal consistency with validity, of norm deficit with needs assessment, of test bias with item bias, of discrimination with utility, and so on. We do sorely need to reach beyond the measurement approach. Because of the inertia of the professional system it may take a more explicit analysis of these errors to do this than it should take, and I hope this will take the lead in this further move towards consciousness raising in the test business.

When we turn to statistics and its interaction with T&M, once again we find the uncommon few who have broken the grip of the rapidly rigidifying discipline—the Mostellers, Lights, and Tukeys who can play the game when they should but expose it as only a game when *that* matters.

The grip of statistics nearly became a death grip five or six years

ago when evaluators brought up in that discipline so dominated the educational R&D area that *only* statistical significance was quoted in the documentation by these evaluators of the results of field tests of educational products. There was *no* mention of raw-score differences. No needs-assessment data to show that the statistical significance was worth an educational damn to anyone in the "treatment" group. We turned up that gem when ETS had the contract for selecting the best of those materials for dissemination funding. But even ETS can slip: if I recall correctly, its evaluation of *Sesame Street* only quoted and certainly highlighted the statistical significance of the intergroup difference. It never focused on the question: Is the actual difference in gains worth the cost of the program to the taxpayer, parent, teacher, and child, all of whom paid a good deal for the treatment? It turned out that the absolute advantage of *Sesame Street* over the standard approaches was tiny, although statistically significant because of the huge number involved.

Well, so much for the ritual hand-biting activity; evaluators have picked up from jurisprudence the adage that it's not enough to bite the hand that feeds you, it's essential that it be *seen* that one bites it.

Moving Closer to Real Decisions

We're by no means out of the shadow of confusing quantity with quality yet—perhaps we never will be. But the evaluator today is less likely than ever before to be deceived by the trappings of quantitative measurement into a conclusion about quality. And she or he is certainly outgrowing the temptation to run back and hide in the shadows while bleating plaintively that "social science is value-free, social science is value-free!" Making evaluative judgments is exactly what science is all about, measurement provides data, but the *use* of measurements requires the *evaluation* of measurements and of much more. What has happened here could be put in this way. As measurement has come to be a smaller part of a larger discipline, as Educational Testing Service has become Educational Testing, Training, and Evaluation Service, as statistics has itself come under evaluation as sturdy or slippery or sloppy, so we have come closer to the world where real decisions have to be made. Made, for example, by judges who spent months reviewing the evidence for special education classes or the effect of racial separate-

ness or the reliability of nonmoney indexes of quality in an educational program, only to find that there are no *useful evaluations* to which they can appeal. As a result of this lack, decisions were made which will cost us billions and may easily have to be reverted. In the law, there are procedures for making decisions in the absence of relevant evidence, as in science. But those procedures would never have to be invoked if we had done our job of performing the evaluations that are relevant to major social programs. We never really wanted to make hard decisions with soft data, so we made the data look hard when it wasn't or developed arguments why *we* shouldn't have to make such decisions at all. The law has never had those options—at least it has had them very rarely since the more creative days of King Solomon. The law *has* to make decisions, however soft the data or the connections between it and the decisions.

As a result, the law has developed methods of reasoning that are substantially different from those of the social sciences; in particular, the logic of argument from precedent and analogy, the logic of prima facie inference, and the burden of proof. It will pay us well to study these and some other, more procedural devices from the law. It is no accident that one of the most interesting models of evaluation today is called the jurisprudential model by Bob Wolff or, in simpler form, the advocate/adversary model by Bob Stake. One way of expressing the latent function of the adversary model is to say that it *legitimates an appropriate range of interpretations*, thereby conveying in a very graphic way the softness of the data. It is the evaluative equivalent of adding the standard deviation to the statistic of the mean—and it is, of course, a standard procedure of the legal system.

In the evaluation sub-area of bias control, to take another example, one line of thought has led to the so-called goal-free procedure in evaluation—one in which the legal idea of justice as blind is extended one step farther than has been traditional in experimental design, where double-blind methodology was picked up from medical research.

And so on. The model of legal process is a fertile one for evaluation. And the model of legal reasoning is no less useful. A few years ago, I had occasion to do a rather detailed comparison between the logical methods (the standards of evidence) in social science and the law for the *Journal of Legal Education*. I concluded that even the striking apparent differences were not real, but I was impressed by

the novel view of scientific reasoning that was forced on one by the detailed comparison. I do not find the New Rhetoric analysis persuasive, but I do believe that there is a need for, and high payoffs from, new models and new study of the logic of the law, as there is for evaluation.

Conclusion

As measurement has matured, as evaluation has developed, we have come to see how much closer *useful* measurement and its useful application to evaluation are moving towards the other practical, decision-oriented disciplines of management, accounting, and law. This may indeed be a development in which *all* the social sciences can share, not just measurement, and so end a too-sterile period in intellectual history. The applied leading the pure, the evaluative and judgmental leading the value-free.

Let it be so!