

From
GATEKEEPER
to **GATEWAY:**
transforming
testing
in America

Report of the
National Commission
on Testing and
Public Policy

FROM GATEKEEPER to GATEWAY:

Transforming Testing in America

**Report of the National Commission
on Testing and Public Policy**

**Copyright 1990 National Commission on Testing
and Public Policy.**

**Published by: National Commission on Testing and
Public Policy, Boston College, Chestnut Hill, MA**

*Design by Joyce Hempstead
Printed by: Knowlton & McLeary*

**NCS General Purpose Answer Sheet No. 6703,
1977 by National Computer Systems, Inc.,
reproduced on the cover of this report with
permission of National Computer Systems, Inc.**

The National Commission on Testing and Public Policy

Members

**Bernard R. Gifford, Commission
Chair Vice-President, Education
Apple Computer, Inc.
Cupertino, CA and
Chancellor's Professor of
Education,
University of California,
Berkeley**

**José A. Cárdenas
Director
Intercultural Development
Research Association
San Antonio, TX**

**Julius L. Chambers
Director-Counsel
NAACP Legal Defense and
Educational Fund, Inc.
New York, NY**

**Frederick R. Chang
Director
Information and Decision
Sciences
Pacific Bell
San Ramon, CA**

**Bill Clinton
Governor, State of Arkansas
Little Rock, AR**

**Badi G. Foster
President
Aetna Institute for Corporate
Education
Hartford, CT**

**George H. Hanford
President Emeritus
The College Board
Demarest, NJ**

**Katharine H. Hanson
Executive Director
Consortium on Financing Higher
Education
Cambridge, MA**

**Antonia Hernandez
President and General Counsel
Mexican American Legal
Defense and Educational Fund
Los Angeles, CA**

**W.W. Herenton
Superintendent
Memphis City Schools
Memphis, TN**

**Francis Keppel
Senior Lecturer
Harvard Graduate School of
Education
Cambridge, MA**

**Robert Linn
Co-Director
The Center for Research on
Evaluation, Standards, and
Student Testing
Professor, University of
Colorado
Boulder, CO**

**Patricia Locke
Executor
International Native American
Language Institute
Mobridge, SD**

**Eldridge W. McMillan
President
Southern Education Foundation
Atlanta, GA**

**Edward Potter
Attorney at Law
McGuiness & Williams
Washington, DC**

**Thomas G. Sticht
President
Applied Behavioral &
Cognitive Sciences, Inc.
San Diego, CA**

**Glenn Watts
President Emeritus
Communications Workers of
America
Chevy Chase, MD**

Staff

George F. Madaus
Executive Director
Boston College
Chestnut Hill, MA

Linda C. Wing
Associate Director
University of California
Berkeley, CA

Walter Haney
Senior Research Fellow
Boston College

Cesar McDowell
Senior Research Fellow
Boston College

Maxwell West
Consultant
Boston College

Robert Lyons
Research Associate
Boston College

Susan Hanson
Administrative Assistant
Boston College

Mary Lou Sumberg
Copy Editor
University of California
Berkeley, CA

Special Advisors

Gloria B. Cabe
Arkansas House of
Representatives
Little Rock, AR

Kaye L. Evleth
Division Chief
Personnel Department
Examining Division
City of Los Angeles
Los Angeles, CA

Albert H. Kauffman
Staff Attorney
Mexican American Legal
Defense and Educational Fund
San Antonio, TX

Damian A. McShane
Associate Professor of
Psychology
Director for the American Indian
Support Project
Utah State University
Logan, UT

W. S. Sellman
Director for Accession Policy for
Force Management Personnel
Office of Assistant Secretary of
Defense
Washington, DC

Peter W. Stanley
Director
Education and Culture Program
Ford Foundation
New York, NY

Kathy VanLaningham
Senior Assistant for Education
Office of the Governor
Little Rock, AR

Lynn Walker
Deputy Director
Human Rights and Social Justice
Program
Ford Foundation
New York, NY

In Memoriam

FRANCIS KEPPEL

1917-1990

The members and staff of the National Commission on Testing and Public Policy respectfully pay tribute to former United States Commissioner of Education, Francis Keppel, a valued member of the Commission until his untimely death. Frank enriched the Commission's deliberations with his wisdom, knowledge, and wry humor. We are grateful for his singular contributions to education and education policy, and admire his real and strategic modesty, his unfailing ability to cut to the heart of issues and to see how things worked, and his gift of bringing people together to get things done.

Table of Contents

Executive Summary

Foreword

1. Introduction

2. Why Testing Must Be Transformed

**3. A Vision of Testing as an Instrument to
Enhance the Development of Human
Talent**

4. A Final Word

Reference Notes

Bibliography

Appendices

National Hearings

Invited Papers

Executive Summary

America must revamp the way it develops and utilizes human talent, and to do that educational and employment testing must be restructured.

America can no longer rely on an abundant, largely unskilled labor supply. Instead, the nation is facing a shrinking entry-level workforce increasingly composed of linguistic, racial, and ethnic minorities, whose talents are often underdeveloped and under-utilized. Yet in a global economy that is becoming more competitive and interdependent, we need more than ever the talents of all our people. Developing that resource is the province of our educational institutions.

From the outset, American education has had the dual goal of creating a skilled workforce and a knowledgeable citizenry. This report deals with the role of testing in pursuit of those goals. We recognize that in the past some tests have been a positive force for numerous individuals and institutions. However, the growing overreliance on testing over the past several decades deprives the nation of much of the talent it needs and sometimes conflicts with the nation's ideals of fairness and equal opportunity.

This report summarizes our findings on the problems of testing

and offers recommendations for its restructuring.

Current testing, predominantly multiple choice in format, is over-relied upon, lacks adequate public accountability, sometimes leads to unfairness in the allocation of opportunities, and too often undermines vital social policies.

Tests may mislead as indicators of performance. Test scores are at best an estimate of someone's knowledge or ability, and can be affected by numerous outside factors. Inevitably, some who could perform successfully will "fail" tests and thus risk being misclassified and erroneously denied opportunity.

Testing can result in unfairness. All tests are to some extent culturally dependent; nor has society yet been able to extend educational opportunities to all — hence the score gap between minority and majority groups. Differences in performance on other indicators such as grades and ratings are generally much smaller than test score differences. Thus when test results alone are used in selection, misclassification falls disproportionately on minority groups.

There is too much educational testing. Mandatory testing consumes some 20 million school days and the equivalent of \$700 to \$900 million in direct and indirect expenditures annually -- an enormous cost and use of classroom time that could be spent on skill development.

Testing practices can undermine social policies. We cannot test our way out of our educational problems; the opposite is true. As teaching turns into test preparation, test results cease to reflect what examinees really know or can do. Thus our fixation on test results deflects attention from fundamental educational problems and so hinders reform.

Tests are subject to insufficient public accountability. Rarely are many important tests and test uses adequately scrutinized; standards for their development and use lack adequate enforcement mechanisms; and truth-in-testing laws exist in only two states. Thus the industry whose products regulate access to opportunities is itself unregulated and unaccountable.

To help promote greater development of the talents of all our people, alternative forms of assessment must be developed and more critically judged and used, so that testing and assessment open gates of opportunity rather than close them off.

This Commission proposes that testing policy and practice be restructured to help people develop their talents and become more

productive, and to help institutions become more productive, accountable, and just. To that end, we offer eight recommendations.

1. Testing policies and practices must be reoriented to promote the development of all human talent.

We must reevaluate how we judge the quality of tests, the names we give them, the ways we report results, and the ways we use them. No testing program should be tolerated that classifies people as unable to learn; potentially negative classification in school or the workplace should be accompanied by learning opportunities.

2. Testing programs should be redirected from overreliance on multiple-choice tests toward alternative forms of assessment.

Important decisions about people and institutions should, where feasible, be based on multiple sources of information, especially direct evidence of actual performance in school and on the job. Thus candidates should supply answers, perform acts, demonstrate skills, create products, and supply portfolios. Previous accomplishments should also be considered.

3. Test scores should be used only when they differentiate on the basis of characteristics relevant to the opportunities being allocated.

For tests to be fair and useful, this differentiation must relate directly to the classifications and decisions to be made. With that aim, evidence should be accumulated to show how well test scores reflect real-life educational or job performance.

4. The more test scores disproportionately deny opportunities to minorities, the greater the need to show that the tests measure characteristics relevant to the opportunities being allocated.

It is essential to evaluate critically the fairness and accuracy of all test-based classifications in terms of the opportunities being allocated, with full awareness of the implications for social groups already disadvantaged. Ensuring equality of educational and employment opportunities is so vital that immediate, but transitional, strategies should be adopted until appropriate forms of assessment can be developed.

5. Test scores are imperfect measures and should not be used alone to make important decisions about individuals, groups, or institutions; in the allocation of opportunities, individuals' past performance and relevant experience must be considered.

Test scores should not be used by themselves to determine kindergarten entry, grade promotion, graduation, or employment opportunities. Furthermore, decision makers' judgments should enter directly into important decisions about people.

6. More efficient and effective assessment strategies are needed to hold institutions accountable.

Assessment of the effectiveness of institutions — e.g., schools and training programs — should differ

from assessment of individuals in order to help them. Large school districts in particular could use sampling techniques to gauge school performance. This would help prevent the distortions caused by using one testing program for both instructional and accountability purposes.

7. The enterprise of testing must be subjected to greater public accountability.

Test quality and use should be subject to some form of independent public scrutiny. Tests would be more accurately labeled, the results constructively reported, and evidence as to what they do and do not measure made more accessible. Scrutiny must include the perspective of groups that have been most adversely affected by testing.

8. Research and development programs must be expanded to create assessments that promote the development of the talents of all our peoples.

Beyond more accurate assessment, we need ways to communicate the uncertainty of all assessment results. In addition, we need to learn how to use multiple sources of information intelligently and sensitively in making decisions. Finally, we need forms of assessment that will prevent unfair classifications.

In conclusion, the Commission recognizes that testing is useful and inevitable: we must know how our institutions are doing, what our children are learning, and who will make the most of opportunities not

available to all. But to direct testing along a more constructive course, we would draw richer, more direct evidence of knowledge and skill from information sources beyond multiple-choice tests. The design of assessments would differ with their purpose — e.g., to inform instruction or to evaluate school performance. And test use would be monitored continuously.

The role of assessment information should be a supportive one; low test scores should never brand anyone as a failure or permanently restrict opportunities.

The shift we envision will be difficult to accomplish. New attitudes and policies cannot guarantee human development. But with resources and national resolve, we can bring testing policies and practices into line with our most important goals and most deeply held convictions.

Foreward

THE NATIONAL COMMISSION on TESTING and PUBLIC POLICY is an interdisciplinary body composed of individuals with expertise, interests, and experience in a wide variety of fields -- education, business, labor, law, assessment and measurement, and manpower development and training. Supported by the Human Rights and Governance Program and the Education and Culture Program of the Ford Foundation, the Commission was formed in 1987, following a preliminary investigation of issues that was organized by Bernard Gifford of the Graduate School of Education at the University of California, Berkeley. Preliminary work also included an invitational conference that established a clear need for a policy examination of the role that tests play in allocating opportunities in American society. The papers presented at that conference were subsequently published in two volumes (see Appendix). In February of 1989 the Center for the Study of Testing, Evaluation, and Educational Policy at the Boston College School of Education assumed the staff work for the Commission.

The Commission's mandate has been:

- To investigate trends, practices, and impacts of the use of standardized test instruments and other forms of assessment

in schools, the workplace, and the military.

- To recommend improvements in testing that would promote the identification and nurturing of talent, especially among racial, ethnic, and linguistic minorities.

Toward that end, over a three-year period, the Commission has heard presentations from a range of experts on a variety of issues related to these two goals. The Commission also invited and reviewed over 50 additional papers (see Appendix), which cover testing among children and adults; among different ethnic, linguistic, and cultural groups; and in the education, employment, and military sectors of American life. Additionally, the Commission convened five public hearings to document the impact of testing on particular population subgroups. This report represents a brief synthesis of what the Commission learned from these many sources. We would urge those who are interested in a fuller treatment of the issues presented here to consult the published volumes of commissioned papers and the companion staff report, which will be released in the fall of 1990.

1. Introduction

THE EMERGING GLOBAL ECONOMY, increasing international competition, and a shrinking labor force make it imperative that the United States fully develop and mobilize the talents of its people. Human talent is the cornerstone of our social and economic future.

In the past, the nation's wealth stemmed from the natural resources that it extracted from the ground. Henceforth, it will be determined by how well we cultivate and use the talent of our people. Primary responsibility for developing this essential resource rests with our educational institutions. President Bush and the nation's governors, recognizing this responsibility, have set national educational goals and called for the full development of everyone in our society.

This report by the National Commission on Testing and Public Policy focuses on the critical role testing and assessment must play, not only in monitoring progress toward meeting the immediate goals established by the President and the governors, but also in the more fundamental tasks of identifying and developing human potential and allocating opportunities from kindergarten through the workplace.

The National Commission on Testing and Public Policy believes that these tasks are essential to the health of our nation. The Commission has been guided by a

vision of a society based on fairness in the allocation of opportunities; respect for the racial, cultural, and linguistic diversity of its people; full implementation of democratic values; and continued global economic leadership. Our democracy is rooted in the consent of the governed — an informed consent requiring a well-educated populace. The Commissioners share the conviction that all our people can learn and contribute to the building of an economically strong and socially responsible nation. After studying how current uses of tests do and do not support these ideals and needs, the Commission recommends a fundamental change in the role of testing in our society that would see testing transformed from a gatekeeper to a gateway of opportunity. Unlocking our greatest national resource requires accurate, appropriate, and responsible assessment instruments, used judiciously and selectively.

America is at a critical stage in its history. In the past, the nation was able to rely upon a largely unskilled and abundant labor supply to fuel the growth of its economy. Tests were most often used to select among plentiful students and workers, and few worried about those rejected. Now, America's entry-level workforce is shrinking and is increasingly composed of members of linguistic, racial, and cultural

minority groups whose talents and capacities have historically often been underdeveloped and undervalued. As the global economy becomes more competitive and interdependent, we will more than ever need the talents of *all* our people. Our challenge will be to create social institutions that

value and benefit from racial, cul-
The Commission recognizes the need for sound, fair, and reasonably efficient mechanisms to help make difficult decisions about individuals and institutions. We affirm society's interests in certifying the competence of in-

The Term “Test” Defined

Technically, a test is a set of questions or situations designed to permit an inference about what an examinee knows and or can do in an area of interest. Most commonly used tests present have select from alternative answers (e.g., multiple-choice tests) or supply oral or written answers (e.g. an essay question, structured interview). A test can also require the examinee to perform an act (e.g., drive a car, read aloud from a book) or produce a product (e.g., compile a portfolio, or write a book report). The term **standardized test** simply means that all examinees are given identical directions, time limits, and questions. A test battery is a set of standardized tests.

From a policy point of view, a test is an **instrument** that yields information that can be used for a variety of purposes; for example, (1) to describe a person’s performance relative to a comparison or norm group (e.g., in a norm group of 100, Mary scored higher than 80 of them on the XYZ math test); (2) to describe what a test taker knows or can do (e.g., Joe can read with understanding material found in a magazine like *Time* or *Newsweek*); (3) to classify a person as knowing or not knowing enough, or is able or unable to perform certain tasks; and (4) to make a decision about the person on the basis of such classifications (e.g., place Eileen in an instructional program, hire Bob but not Bill). Test results also can be aggregated to describe, classify, or make decisions about groups of persons or institutions.

Whether from a technical or policy perspective, the key question to ask about testing is, “How accurate are the inferences, descriptions, classifications and decisions made about individuals or institutions on the basis of test performance?” One’s answer to this question goes not only to the heart of the technical question of test validity but also to the policy dimension of fairness in the allocation of opportunities.

How the Commission Use The Word “Test”

The term **test** is generic and encompasses many forms and techniques, such as multiple-choice questions, essays, structured interviews, and a variety of products and performances.

The Commission is most concerned, however, with overreliance on the form of testing commonly used to allocate initial opportunities in education, employment, and the military: group-administered, paper-and-pencil, multiple choice tests. Throughout this report, the term **test** is used only to refer to these instruments. The term **assessment** is used to denote a broader array of devices designed to show what a person knows or can do. This term is always used in referring (1) to instruments and methods that require a test taker to supply an answer, a product, or a performance; and (2) to the use of diverse instruments and methods for arriving at a description, classification, or decision. Assessment may include use of multiple-choice tests along with other tests.

Defining “Minorities” and “Diversity”

Throughout this report the term **minority** refers to the following groups and the individuals who comprise them:

- Black Americans
- American Indian Nations and their citizens, and Alaska Natives
- Asian Americans and Pacific Islanders, including Native Hawaiians
- Hispanic Americans, including Cubans, Mexican-Americans, Latin Americans, and Puerto Ricans.

When the Commission speaks of **diversity**, we refer to the mix of all these groups, and to all women, majority and minority, as well. Women are included in this usage because, while not a numerical minority, they are subject to the same protections against discrimination under civil rights law.

dividuals and holding institutions accountable. Assessment has an important role to play in helping society to realize these interests. The Commission recognizes that some tests have been a positive force for numerous individuals and institutions in both education and employment. A significant segment of society, however, has experienced and continues to experience testing as a hostile gatekeeper: although tests open gates of opportunity for some, they slam them shut for many others.

The focus of this report is on the use

of tests and assessment to open the gates for America’s diverse people. In this report, we summarize our findings on the problems of testing and offer recommendations for the restructuring of testing to expand opportunities and promote the development and utilization of the talents of all our diverse peoples. Achieving this goal will require bold rethinking of the ways we conduct large-scale testing programs and use the results. It will also require the development and use of fairer and more appropriate methods of assessment.

2. Why Testing Must Be Transformed

THE COMMISSION FINDS THAT overreliance on testing to often carries an exorbitant price. Our central finding is that many current practices in educational and employment testing stand in the way of efforts to identify and develop talent, and to improve the functioning of key social institutions.

Over the past three decades tests have come to be used as administratively convenient and inexpensive tools to help solve an array of troubling problems in education and the workplace. Only gradually are the unintended negative results of using tests as instruments of policy becoming clear. The story of Antonia Gonzalez vividly illustrates such a situation.

Antonia, an A student in high school, dropped out in her senior year to get married. Years later, at age 40, she completed her high school equivalency requirements, enrolled in a junior college, and finished two years with a 3.4 grade-point average. She subsequently transferred to a university to pursue a degree in education, earning B grades in her first two upper-level education courses. Antonia then encountered a roadblock: she failed the “rising juniors” test, recently required by her state for admission to an approved program of undergraduate teacher education.

Antonia became a statistic, part of the 66 percent Hispanic failure rate on the “rising juniors” test. Although she passed the mathematics section of the test on her first attempt, she failed the reading and writing sections. A year later, after

taking every test preparation course she could, she passed the writing test but failed the reading test by a single point. In her third attempt she again failed the reading test by one point and was prohibited from taking further upper-level education courses until she passed.

Antonia’s initial opportunity to prepare for a career in teaching was denied on the basis of a single point on a single test despite evidence of her previous success in upper-level education courses. Her employment record since leaving the university without a teaching credential suggests that she could be a valuable asset to the teaching profession in a state where so many students, like herself, are native Spanish speakers. She went on to tutor in a home for disturbed boys. Experienced educators judged her teaching there to be outstanding.

Why was Antonia barred by a test from a profession for which she showed promise and which desperately needs teachers from her ethnic and linguistic background? Many approach this question by asking another: “Why did she fail the mandated test?” There are several possible answers to this question. Antonia was an older student. English was not her first language. Measurement error and culturally and linguistically biased test questions may have negatively influenced her scores. All these factors may explain why Antonia “failed” to hurdle the wall of her state’s mandated test.

But to focus on why Antonia did not get higher scores on the test is to miss the broader and far more vital answer to the first question. Antonia did not fail the test so much as the testing policy of her state failed both her and the educational needs of her community. The policy ignored the fact that since any test is a fallible instrument, some scores will always turn out to be lower than they should. More importantly, when a test score *alone* is used as a bar to educational or employment opportunity, some capable people — especially people of ethnic and language minority backgrounds — are inevitably misclassified. When clear and direct evidence of a person’s competence is available — in Antonia’s case, both satisfactory performance in college and success on the job — a test score should not be allowed to present an absolute bar to opportunity. The way the test was used was thus not only illogical; it was unfair. It misdirected Antonia’s education by wasting her time and money in test preparation courses. Further, it undermined her state’s efforts to encourage minority students to enter the field of teaching.

Three years of study have convinced this Commission that stories such as Antonia’s are common to many educational and employment testing programs, especially those that use test scores *alone* to determine initial opportunities.

Antonia and thousands like her bear the often hidden negative consequences of the exclusive use of testing as an instrument of social policy. While assessment as an instrument of social policy is both necessary and inevitable, the Commission focuses on the following limitations of testing:

- Tests are imperfect and therefore potentially misleading as measures of individual performance in education and employment.
- Some test uses result in unfair treatment of individuals and groups.
- Students are subjected to too much testing in the nation’s schools.
- Some testing practices in both education and employment undermine important social policies and institutions intended to develop or utilize human talent.
- Tests have become instruments of public policy without sufficient public accountability.

These concerns form the basis for the Commission’s recommendation that the role of testing in the United States be transformed.**

* Although Commissioner Linn supports the recommendations of the report he believes that the findings give too little attention to the benefits of many current testing programs. He points out that certainly tests are imperfect as either indicators of current achievement levels or predictors of future accomplishments, but in many instances they are superior to available alternatives.

The Consequences Associated With a Test's Name

Names Given to tests, such as *intelligence*, *ability*, *competence*, *aptitude*, *readiness*, or *honesty*, often encourage mistaken beliefs about what they can do. People too often take such labels literally. Further, diverse emotional and cultural connotations and images associated with a test's name are often transferred to one's interpretation of test performance. A test's name can profoundly affect attitudes toward its use. Many who oppose the use of an "intelligence test" to determine kindergarten entrance unwittingly accept the practice when an almost identical test is called a "readiness test." The Commission finds that, whatever its name, using any paper-and-pencil test to exclude children from kindergarten or first grade is intolerable.

Tests are imperfect and therefore potentially misleading as measures of individual performance in education and employment.

Paper-and-pencil tests are often viewed as precise measures of attributes ranging from general "intelligence" or scholastic "ability" to vocational "aptitude" and "trainability." These views invite misinterpretation of what the tests can do. No test is a *direct* measure of aptitude or innate ability. Tests sample only a small portion of what someone knows or can do at a particular time — the time of the test.

Despite their mystique of numerical precision, test scores are at best only an estimate of what someone actually knows or can do. Scores can be influenced by many factors irrelevant to what the test is measuring that cause persons with equal levels of skill or knowledge to earn quite different scores. Such factors include:

- Test anxiety
- Test sophistication
- Test-taking habits
- Test directions

- Attitudes toward tests
- Noise level of the testing environment
- Language of the test
- Native language of the test taker
- Cultural background of the test-taker
- Special coaching
- And even not having eaten breakfast.

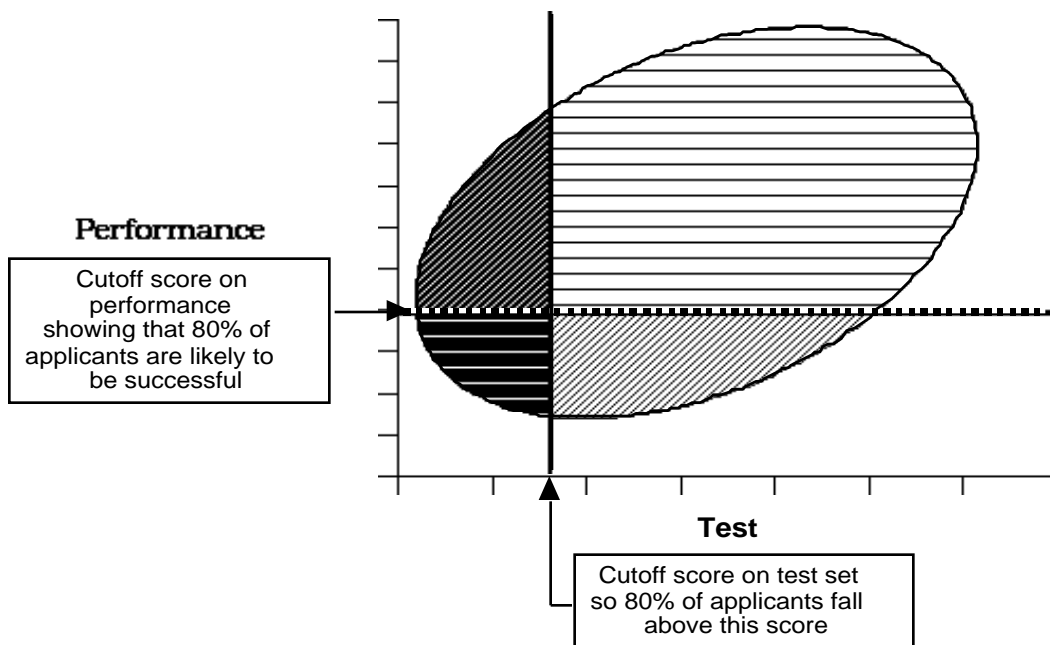
More than a half-century of studies of both college admissions and employment tests illustrates that neither scores on well-developed tests nor any other single indicator can precisely predict how school or workplace performance will differ from one person to the next. For example, numerous studies have shown that scores on college admissions tests have only a limited relationship with how students perform in college. Other studies have found that many employment tests predict job performance even less well. This is because factors important to doing well in college and on the job -- such as going to others for help when needed, study or work habits, cooperation with other people, and a host of institutional support

factors — are not reflected in test scores.

The general problem of the fallibility of test scores in predicting real-life performance has long been known to specialists in testing. However, because test scores are increasingly being used as absolute arbiters in the allocation of opportuni-

ties, the consequences associated with the problem are becoming more serious and conspicuous. Given what is known about the imperfect power of tests to predict actual performance in school or on the job, it is inevitable that many people who could perform successfully will “fail” the test.

Figure 1: Imperfect Classification Based on Test Scores and Performance*



Classification Key

	Above cutoff score on both test and performance	Correctly classified
	Below cutoff score on test but likely to succeed according to performance indicator	Misclassified
	Below cutoff score on both test and performance	Correctly classified
	Above cutoff score on test but unlikely to succeed according to performance indicator	Misclassified

*Note: the data illustrated represent test scores with typical power to predict performance (i.e., a correlation of 0.35 between test scores and indicators of performance).

For example, consider the case where 800 people out of a group of 1000 could perform successfully on the job or in school. If a test with the typical power to predict performance is used to classify them into two groups – likely or unlikely to succeed – about 66 percent of the 200 who “fail” the test barrier could actually perform successfully; and of the 800 who “pass” the test about 17 percent would be unlikely to perform successfully. This classification situation is illustrated in Figure 1.

The problem of mislabeling people when classifying them on the basis of test scores is well known both in history and in theory. But in many testing programs over the last several decades this fundamental problem has been obscured by technical procedures developed for setting cutoff scores. These procedures involve using the judgments of persons who are considered “experts” in the subject matter of the tests (for example, experienced teachers for teacher certification tests, or law enforcement officers for police promotion tests) about the importance of each test question to the intended classification. Such analyses often give a scientific aura to the process of setting cutoff scores. However, research has shown that the cutoff score arrived at for a given test may vary substantially depending on the method used to elicit these judgments and the composition of the expert review panel. Moreover, though use of these procedures may indicate one cutoff score, political and economic considerations can lead to the eventual choice of a different cutoff score. For example, personnel

administrators may raise cutoff scores on employment tests to prevent too many interviews; and school board members sometimes lower cut-scores on minimum competency tests because the fiscal costs associated with retention in grade or remedial programs are too high.

It is not surprising that economics and politics influence testing programs. What is important to note is that, however a cut-score — say 70 out of 100 questions right — is determined, it is arbitrary; it has no scientific basis analogous to that underlying 32 degrees Fahrenheit on the temperature scale, and no physical basis analogous to a cutoff score of 20/250 on the scale of visual acuity used to define legal blindness. Other cut-scores used in public policy such as speed limits, air pollution standards, and tax brackets are also arbitrary. However, they do not go into effect without extensive public hearings, debate, and legislative action.

Because cut-scores used in testing programs are arbitrary, many classifications based on a single cut-score are also arbitrary. The accuracy of classifications should be examined by asking,

“Of those who *can perform satisfactorily*, how many *fall below* the cut-score on the test?”; or

“Of those who *cannot perform satisfactorily*, how many *score above* the cut-score on the test?”

The answers to these questions become increasingly important as the number of new entrants to the labor market continues to decline. How-

ever, the questions are not often asked, much less answered. Many classifications based on a test score exclude people from jobs at which, given the initial opportunity, they would have been successful. This phenomenon was graphically illustrated when an error occurred in the calibration of the 1976 version of the Armed Services Vocational Aptitude Battery (ASVAB). The ASVAB, part of the Department of Defense's large testing program, is used to screen potential entrants into the armed services. As a result of the calibration error, more than 300,000 recruits were admitted into the military who normally would have been rejected because of low scores. Several studies

show that as a group these enlistees performed only somewhat less well than those who actually passed the ASVAB. Numerous enlistees fared as well or better

The ASVAB story shows that whenever people are classified on the basis of cutoff scores on tests, misclassifications are bound to occur. Some who score below the cutoff score could perform satisfactorily in school or on the job, and some who "pass" cannot perform satisfactorily. The solution to this problem is not to avoid classifying people: such classifications are essential and inevitable in modern society. Instead it is to avoid classifying solely on the basis of one imperfect instrument.

Cut-Scores and Political Expedience

In the case of *Richardson v. Lamar County Board of Education*, plaintiff Richardson contested the Board's decision not to renew her teaching contract because she failed to pass the Alabama Initial Teacher Certification Tests. In ruling for the plaintiff, Judge Myron Thompson concluded that the method of determining the cut-scores was "so riddled with error that it can only be characterized as capricious and arbitrary."

The cut scores...were so astoundingly high that they signaled...an absence of correlation to minimum competence. For example, of the more than 500 teachers who took the first administration of the core examination, none would have passed.

Faced with this problem, the test developer made various mathematical "adjustments" to the original cut score...The State Department of Education was then given option of lowering the cut scores...which clearly, even after the various adjustments...were not measuring competence.....Instead of challenging what the developer had done, the state simply lowered the cut score...to arrive at a "politically" acceptable pass rate.

The State Board of Education and the test developer in effect abandoned their cut-score methodology, with the result that arbitrariness, and not competence, became the touchstone for standard setting.

Richardson v. Lamar County Board of Education, civil action no. 87-T-568-N (Middle District of Alabama, No. Div.).

Some test uses result in unfair treatment of individuals and groups.

Some misclassification of people is inevitable because all of our methods of predicting human behavior are imperfect. But when test results are used *independently* of other relevant information about people to classify them in ways that affect their opportunities, misclassification can amount to substantive unfairness. For example, when “readiness” tests are used to determine whether children may enter kindergarten or first grade, up to a third of the children who fall below the cut-score may be misclassified and incorrectly denied access to services. And these are the very children who might benefit most from being in school.

Moreover, when test results alone are used to allocate opportunities, unfairness often falls disproportionately on certain ethnic, linguistic, and cultural minority groups. By unfairness the Commission is referring to the fact that disproportionately fewer minorities who could perform successfully had they been given the opportunity will be selected. While the Commission recognizes that there are other perspectives on fair test use, the performance perspective on fairness will be increasingly important as available human talent is in short supply in the next century. Unfairness from the performance perspective (and from some other models of fairness in testing) results from the following combination of factors:

- The fallibility of test scores for predicting educational or job performance
- Relatively large test score differences between majority and minority candidates
- Relatively smaller differences between majority and minority candidates on the actual performance that test scores may be used to predict.

The fallibility of test scores for predicting educational and job performance was discussed in the previous section. Here we explain how the two other factors conspire with test score fallibility to make allocation of opportunities based on test scores unfair for minority individuals.

As has been widely recognized, there are fairly consistent differences in average test scores of minority and non-minority groups. Differences between average scores of black and white Americans have been found to be fairly substantial in studies of both education and employment tests, with typical findings showing that only around 16 percent of black Americans score above the median score of white Americans. Black Americans in recent years have scored on average about 100 points below white Americans on the Scholastic Aptitude Test, for example.

Not only black Americans, but also Hispanics, American Indians, native Pacific Islanders, some Asian Americans, and other minorities tend to score significantly lower than their majority peers on many tests. These group differences are large and fairly consistent on tests ranging from

kindergarten entry tests to tests used in elementary and secondary schools, and from college and post-college admissions tests to vocational aptitude and employment tests. Although specific results vary from place to place and test to test, the overall finding is that on average, racial, cultural, and linguistic minority group members score below majority examinees. Women as a group also score lower than men on certain important tests such as college entrance exams.

There are two broad reasons for group differences in test performance. First, tests, as well as other methods of predicting behavior, are culturally bound and almost always reflect the dominant or “national” culture in both form and content. Research shows that the way test content is oriented — toward the topics and culture of one group as opposed to another — can significantly affect test scores. A recent report of the National Research Council pointed out, for example, that “even carefully

Differential Test Performance

The Commission has found stable group differences in performance for each of the subgroups the subgroups targeted by our investigation on most major types of high-stakes tests. A sample of those findings follows.

Minimum competency tests are used to determine grade promotions and high school graduation. About 40 percent of American Indians fail state-mandated competency tests, versus 20 percent of their non-Indian peers.

College admissions tests are used to inform admission and to determine scholarship eligibility. In 1988, women averaged 56 points lower than men on the Scholastic Aptitude Test. Relative to Whites, black Americans averaged 92 points lower, Puerto Ricans 90 points, Mexican Americans 63 points, and Asian Americans 37 points.

Rising junior tests are required for promotion to upper-level professional preparation programs. On the most recent administration of the College Level Academic Skills Test, of the students taking the examination for the first time, only 64 percent of the black Americans and 71 percent of the Hispanics passed all four subtests, while 92 percent of the Whites did so.

Graduate admissions tests are used to screen applicants for post-graduate professional programs. In 1988, Whites outscored all other ethnic groups taking the Graduate Record Examination on the verbal and analytic subtests. Only Asian Americans outperformed Whites on any subtest averaging 63 points higher on the quantitative section of the exam.

Licensing and certification examinations are often the final hurdles for those who have completed all other educational requirements to practice the professions of their choice. During the 1984-85 school year, 45 percent of the Asian and Pacific Americans taking California’s teacher certification exam failed. Many passed the required mathematics test, but failed the English reading and writing sections

Sources: See Notes

designed test instruments may include some degree of cultural bias that artificially lowers the tested performance of [black Americans] relative to [white Americans].”

The second cause of differential group performance on tests is economic and educational. Many ethnic, linguistic, and cultural minorities suffer enormous economic disadvantages — such as lower average incomes, higher rates of unemployment, and jobs with significantly lower occupational status than majority group members — as well as serious educational disadvantages.

Group differences in test scores are consistent with these broad social inequalities. The economic status of individuals’ families is associated with test scores. Students with parents who have more years of schooling tend to have higher test scores than those whose parents have fewer. High school dropouts tend to have lower test scores than high school graduates. Thus, in part, group differences in test scores simply reflect social realities. And the use of test scores in isolation to allocate opportunities compounds these inequalities

Despite large test score differences between minorities and non-minorities, indicators of actual performance in education (such as grade point averages) or on the job (such as supervisor evaluations) do not show similarly large group differences. For example, though black Americans applying to college do have high school grade point averages that are somewhat lower than those of white Americans, on average the gap is not as large as the college admissions test score gap.

The crucial point is that the limited power of tests to predict success in either schools or the workplace, coupled with test score gaps larger than performance gaps, means that using test results alone to classify people will result in higher rates of misclassification for lower scorers, particularly for minority groups who tend to score poorly on the tests. This is true regardless of whether the test score gap between minorities and non-minorities arises from economic and educational disadvantages, from biases in the tests, or from an interaction of these factors.

Culture and Test Performance

A young Ojibwa (Chippewa) student was tested by several professionals and classified as having certain learning and behavioral problems. In part, this classification was based upon his staring vacantly into space, completing tasks very slowly, and giving "non-reality based" responses to questions. As it turned out, the boy had a special relationship with his traditional Ojibwa grandfather, who encouraged his dreaming whether it occurred by day or by night and often discussed the nature of dreams with him.

In Ojibwa thought and language, **ga-na-wa-bun-daw-ming**, which means seeing without feeling (objectivity) carries less value than **mu-zhi-tum-ing**, which means feeling what you do not see (subjectivity). The Pascua Yaqui, Northern Ute, and Red Lake Chippewa Nations have enacted education codes that include the ability to daydream as one of the criteria for identifying gifted and talented students.

McShane, D. (1989, October). Testimony presented at a hearing the The Effects of Testing on American Indians, co-sponsored by National Commission on Testing and Public Policy and the Native American Scholarship, Inc. Albuquerque: University of New Mexico.

An example of this phenomenon was presented in a recent report of the National Research Council concerning fairness in employment testing. The example was drawn from a study of working carpenters, including both black and white Americans. The carpenters were all given a Department of Labor employment test and rated as to job performance by their supervisors. The results showed that if the test had been used with a cut-score to differentiate carpenters whose job performance was good from those whose job performance was poor, 15 percent of the white carpenters and 50 percent of the black carpenters rated as good performers by their supervisors would have been screened out. As the National Research Council Committee pointed out, these results “say that good black workers will be disproportionately screened out in a test-based referral system, and unsatisfactory white workers disproportionately screened in. The test is performance biased against black workers.”

Fortunately, evidence has already shown the superiority of various forms of assessment over multiple-choice tests for ethnic, cultural, and language minorities. Existing alternatives to current multiple-choice tests have demonstrated significantly less adverse impact on minorities while providing information that is at least as helpful to decision makers. Using students' high school records in making decisions about college admission, for example, predicts their future performance at least as well as admissions tests, and the groups differ

far less. Both used together are better predictors than either used alone. Similarly, alternative forms of employment assessment -- such as trainability tests, samples of work performance, biographical data, and assessment centers -- have proved useful in predicting worker performance with less negative impact on the employment opportunities of minorities than that associated with existing cognitive paper-and-pencil tests.

Students are subjected to too much standardized testing in the nation's schools.

The Commission conservatively estimates that each year the equivalent of over 20 million school days is given over to students in the nation's elementary and secondary schools *simply taking* standardized tests! And this figure does *not* include time devoted to test preparation. Far too much valuable teacher and student time is consumed by mandated state and district-level testing.

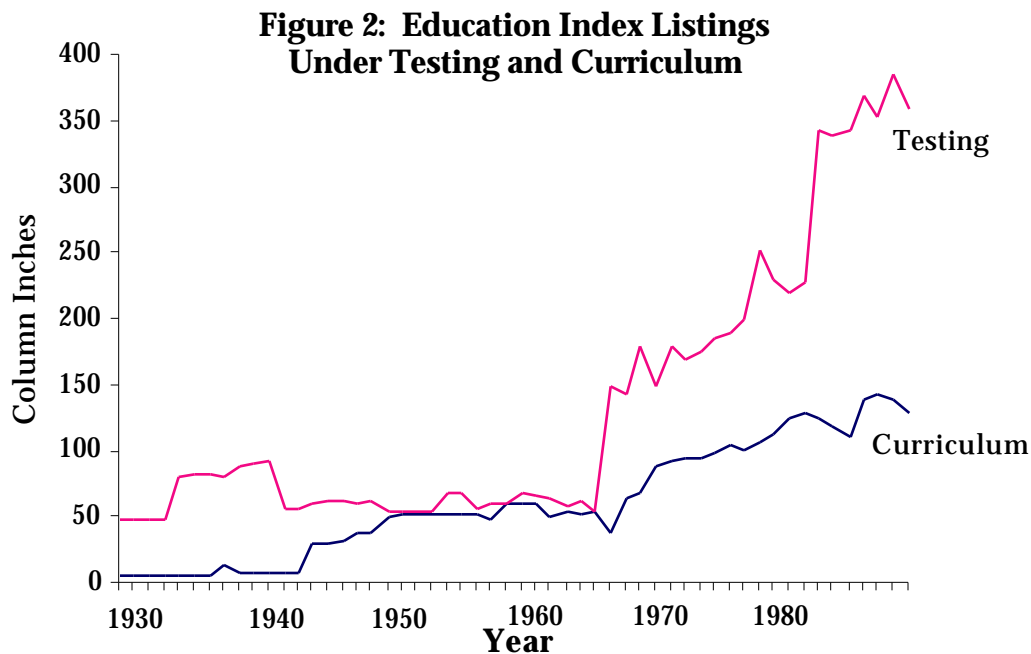
The Commission finds that testing primary school children for entry to or exit from a grade is poor education practice. Nonetheless, pre-kindergarten tests are mandated in more than 16 states, widely used in seven states, and known to be used at the district level in more than 37 states. Kindergarten exit/first-grade entrance tests are used in at least five states and known to exist at the district level in an additional 37. In some school districts as many 60 percent of the kindergartners are judged to be “unready” for first grade because of their scores on a “readiness” test. Achievement testing is

required for first graders in nine states; for second graders in nine states; and for third graders in 27 states.

Overall, the Commission estimates that each year elementary and secondary students take 127 million separate tests as part of standardized test batteries mandated by states and districts. (A test battery may include separate tests that measure such things as reading, mathematics, science, and study skills.) At some grade levels a student may have to take as many as seven to 12 such tests a year. Testing is generally heavier for students in special education or bilingual programs.

The following three indicators illustrate the extensive growth in testing.

- In the *Education Index*, a yearly listing of all educational literature, entries under the categories of curriculum and testing show that increasingly more attention in the literature is given to testing than to what should be taught. While the average annual number of column inches devoted to curriculum issues barely doubled from the 1930s to the 1980s, the number devoted to testing citations increased 35-fold.



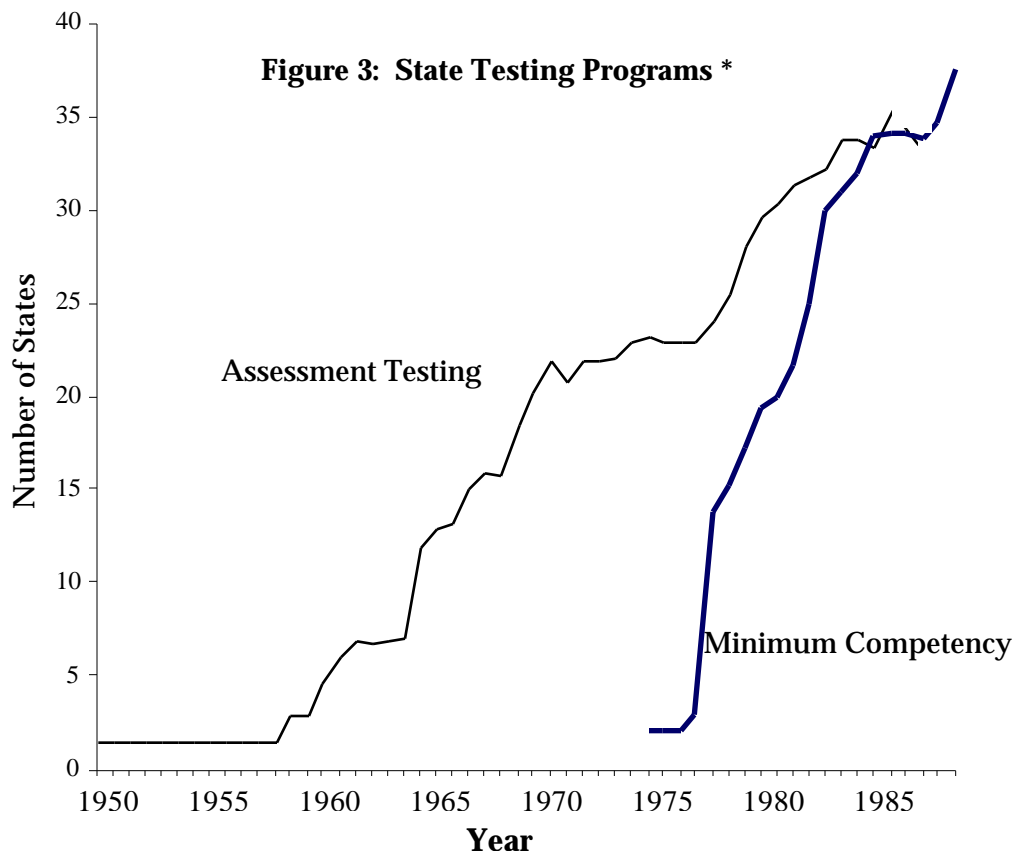
Source: *Education Index*, 1930-1988

- From 1972 through 1985 the number of state testing programs grew from one to 34. By 1989 every state had a mandated testing program of some sort.
- From 1955 to 1986 the *reported* dollar volume of sales of tests and testing services at the elementary and secondary level (referred to in the industry as the Elhi market) grew by almost 400 percent. Reported sales (in 1988 dollars) rose from less than \$30 million in 1955 to over \$100 million by 1986. *Actual* Elhi-related revenues from sales of tests and related services may be four or five times higher, more on the order of half a billion dollars a year.

Since the 1970s, the dramatic increase in test sales and use has been coupled with the disturbing trend of relying more and more on test results to make critical decisions about children, such as:

- Entry to and exit from kindergarten
- Promotion from grade to grade
- Placement in remedial programs
- Graduation from high school.

There has also been a dramatic increase in the use of students' scores to hold school systems, administrators, and teachers accountable. Thus, not only has the volume of testing increased, but testing now looms more ominously in the lives of many educators and children, influencing what they teach and how, and what they learn and how.



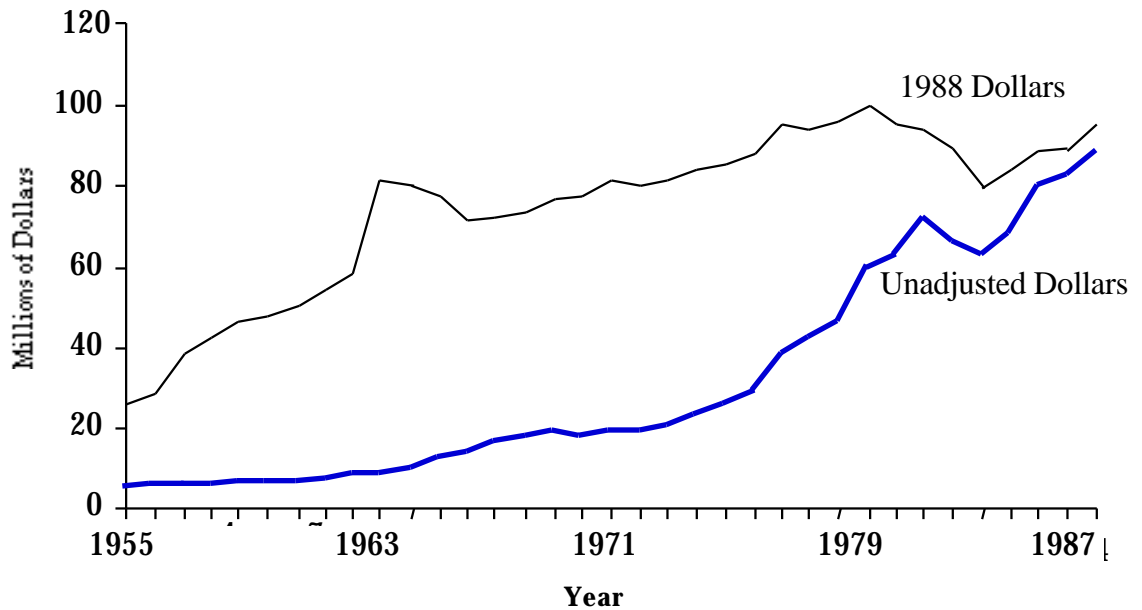
Source: Office of Technology Assessment, US Congress, 1987

*Number of states implementing minimum competency and assessment programs

The figures on the extent and growth of testing mask considerable monetary costs. The Commission estimates that:

- Direct costs to taxpayers of purchasing and scoring state and local tests range from \$70 million to \$107 million annually.
- Indirect costs of this testing, in terms of teacher and administrator time simply spent administering the tests, is in the range of \$268 million to \$421 million, or three to four times the direct costs.
- Indirect costs associated with teacher time given over to test preparation conservatively adds \$387 million to the cost of education.
- Overall, the direct costs to taxpayers for state and local testing plus indirect teacher costs total between \$725 million and \$915 million annually. These costs are well over the recently proposed increase in the Head Start budget.

Figure 4: Reported Dollar Volume of Sales of Tests and Test Services at the Elementary and Secondary Level



Source: *The Bowker Annual*, 1970-1987

Further, time spent teaching to the test must be viewed as an opportunity cost. That is, the classroom time taken to prepare students specifically for statewide and school-district tests could be put to any number of alternative instructional uses. The effects of mandated tests on mathematics instruction illustrate an important opportunity cost borne by the nation. A national study conducted by Thomas Romberg and colleagues of the National Center for Research in Mathematical Sciences Education recently concluded that math teachers who spend the bulk of their effort teaching to the yardstick of present standardized tests do so at a tremendous sacrifice. They forego the intellectual engagement students need to develop the kinds of mathematical abilities recommended by the National Council of Teachers of Mathematics. The nation can no longer afford these opportunity costs.

Our purpose is not to set definitive fiscal or opportunity costs, but rather to put them into a perspective that cannot be ignored. Our cost estimates might be construed by some as legitimate educational costs, *except* for two intolerable facts:

- Dissatisfaction with the education of the nation's students has been increasing even as testing in the schools has become more prominent. This observation is not meant to imply that increased testing has caused deterioration in educational quality, but it does indicate that we cannot test our way out of our educational problems.
- Too many tests, emphasizing lower-order thinking skills, are

devouring valuable instruction and learning time, at the expense of developing the transferable and flexible learning and problem-solving skills the nation wants for its graduates. This is particularly unfortunate because, as the complexity and pace of change in society increases, the need for problem-solving skills becomes critical.

Policy makers must scrutinize the fiscal and opportunity costs of present testing programs and compare them with their tangible benefits. They need to ask questions such as:

- Is this the best that educators can do with the amounts of money and time now devoted to testing programs?
- Are teachers empowered by the test information they receive?
- Do students benefit from the standardized testing they take?
- Are instruction and learning better for all the testing that takes place?

The Commission's answer to these questions is no. The nation cannot continue to devote such enormous human resources (in student, teacher, and administrator time) to the testing programs we now have. Some of these fiscal and human resources must be diverted toward the development and use of other assessment techniques that will directly benefit both teaching and learning.

Some testing practices in both education and employment are undermining important social policies and institutions intended to develop or utilize human talent.

The increase in the sheer volume of testing and the linking of test performance to critical decisions conspire to give us false information about the learning of our students and the occupational capabilities of our young people. Further, the current infatuation with test results as the principal yardstick of educational quality deflects serious deliberation from the underlying problems in the schools, hindering efforts at school reform.

The fixation on test results and the importance attributed to them compromise first the educational processes they were meant to assist, and eventually the classifications and decisions made on the basis of the test scores. As teachers shift their instruction to mirror the form and content of the tests, test results eventually cease to reflect what students really know or can do. In a now famous 1987 report, John Cannell, a West Virginian pediatrician, showed that the vast majority of school districts and all states were scoring above average on nationally normed standardized tests -- a statistically miraculous happening. Cannell's study -- nicknamed the "Lake Wobegon Report," after Garrison Keillor's mythical town where "the women are strong, the men are good-looking, and all the children are above average" -- concluded that these results were implausible and misleading. That conclusion was strongly attacked by some test publishers and other critics. Nonetheless, a study funded by the U.S. Department of Education, and conducted by Robert Linn and colleagues at the Center for

Research on Evaluation, Standards, and Student Testing, confirmed Cannell's basic finding that test results across the nation are inflated. Another study conducted by Lorrie Shepard of the same Center found that the conditions producing that inflation -- such as important rewards and sanctions directly linked to test results, efforts to align curricula with the tests, and direct teaching to the tests -- exist in virtually all the states.

A number of other recent studies reveal a wide and troubling range of test preparation practices in schools, from the legitimate fostering of testwiseness, through aligning instruction with test content, teaching to the test, teaching the test, testing children with a test meant for those at a lower grade, and exempting low-achieving children from taking the tests, to outright fraud and cheating. Some states provide test preparation material, and the sale of a host of commercial test preparation materials has become a significant spin-off of the testing industry. Since standardized tests have become the principal yardsticks of educational quality, most of these practices tend to undermine efforts at educational reform by providing educators, policy makers, and the public with a misleading picture of the schools.

The pressure to improve test scores has trickled down to kindergarten and first grade. Many early childhood programs and teachers emphasize rote academic learning at the expense of exploratory play and social learning. In the upper grades the pressure to show improved reading and math scores tends to turn teaching in these critical subjects into test preparation. For example, instead of reading books, students in many class-

rooms read isolated paragraphs and practice answering multiple-choice questions about them. Thus, the importance attached to test scores is driving schools and teachers away from instructional practices that would help to produce critical thinkers and active learners.

Placing excessive reliance on test results diverts attention from deficiencies in the system and saps the energy and resources needed to address them. Moreover, until underlying educational problems are addressed, employers' lack of trust in the skill levels of high school graduates will continue to spawn proposals such as the recent one by the American Business Conference to create a new national standardized test of basic and academic skills for high school graduates. Although the Commission understands the motive for this proposal to assess performance, we believe it to be unsound and delusive. First, we need much more thought and discussion about the implications, costs, and benefits of *any* type of high school exit test. Second, like many

state and district educational tests such as an important national test would likely lead to outright test preparation, so that eventually the test results would not tell employers what they want to know.

By contrast, effective school reform would make the need for such a measure moot and would greatly reduce employers' dependence on tests for decisions about initial employment. The answer to producing graduates with marketable, transferable skills and the ability to continue to learn throughout life lies in better education, not in more machine-scored, standardized testing.

Testing programs can also have unanticipated negative effects on government policies to help the unemployed. The National Research Council, in its review of the General Aptitude Test Battery (GATB), pointed out how the very existence of a testing program can serve as a powerful screen within populations who expect to perform poorly. For example, when a new GATB screening test was instituted on a trial basis in one New

How Testing Influences Instruction

In a paper prepared for the National Commission on Testing and Public Policy Edward Haertel describes how testing can undermine effective instructional practice.

There has been a subtle shift, especially at the primary and upper elementary levels, toward instructional activities resembling objective test formats. This shift reflects not only the importance of good student test performance, but also the effect of specifying intended learning outcomes in the language of measurable, behavioral objectives. Classroom discussion, simulations and small-group activities, and extended writing opportunities will do less to improve test scores than will worksheets requiring students to answer brief, isolated questions by filling in blanks or selecting among fixed choices. Of course, teachers use such activities not primarily to improve test scores, but to foster student learning. Nonetheless, their use of worksheets and practice tests is likely to increase if they accept objective tests as valid measures of most important learning outcomes. If teachers believe that the goal of schooling is to shape a certain behavioral repertoire, then worksheets and practice tests are the kinds of instructional activities that make sense.

Haertel, E (1989). Student achievement tests as tools of educational policy: Practices and consequences. In B.R. Gifford (Ed.), *Test policy and test performance: Education, language, and culture* (pp13-14) Boston Kluwer

Jersey Job Service Office, the experiment had to be abandoned because many potential applicants, especially minorities, stopped using the placement and referral services.

Tests have become instruments of public policy without sufficient public accountability.

Today those who take and use many tests have less consumer protection than those who buy a toy, a toaster, or a plane ticket. Rarely is an important test or its use subject to formal, systematic, independent professional scrutiny or audit. Civil servants who contract to have a test built, or who purchase commercial tests in education, have only the testing companies' assurances that their product is technically sound and appropriate for its stated purpose. Further, those who have no choice but to take a particular test — often having to pay to take it — have inadequate protection against either a faulty instrument or the misuse of a well-constructed one. Although professional standards for test development and use in education and employment have been formulated by the American Psychological Association, the American Educational Research Association, and the National Council for Measurement in Education, they lack any effective enforcement mechanism.

Another attempt to increase accountability in educational testing has been the passage in two states of “truth-in-testing” laws that allow test takers to see their answer sheets and the key of correct answers. This limited information, usually combined with limited resources and expertise in

testing, makes it difficult for examinees to evaluate the soundness of decisions based on tests. Furthermore, truth-in-testing laws do not enable examinees to evaluate the adequacy of a test as a measure of what it purports to measure, or to question how the test is used.

In employment testing, a set of test standards with an enforcement mechanism has existed for two decades. Title VII of the Civil Rights Act of 1964 outlawed discrimination in employment with respect to race, color, religion, sex, and national origin. The Act also established the Equal Employment Opportunity Commission (EEOC), which issues guidelines for the legality of employment testing procedures. Any test that adversely affects the hiring, promotion, or other employment opportunities of anyone protected under the Civil Rights Act constitutes illegal discrimination unless it can be shown that the test is job-related. Because Title VII covers only employment testing, no other testing can be challenged under it.

The most common way to challenge important tests is through the courts, either under a Title VII claim or through other avenues. Whatever the mechanism, litigation has a number of serious drawbacks. First, in education, when a legal challenge is mounted, the test questions often can be seen only by expert witnesses, and testimony about their quality is given in secret. Many problems associated with such publicly funded tests thus may not become public, particularly if the court challenge is unsuccessful. Second, under present laws a test that is equally harmful to

both majority and minority candidates is not easily subject to challenge. Finally, even when court challenges succeed and compensatory damages are awarded, the cases often drag on so long that opportunities for work and learning may be denied claimants for years.

The lack of an audit or regulatory agency, the absence of mechanisms to interpret and enforce existing professional test standards uniformly, and the limitations of court challenges mean that the industry that develops the products used to regulate access to opportunities, and to hold individuals and institutions accountable, is itself largely unregulated and unaccountable. Government-sponsored testing in our society is too important, and the consequences to test takers too serious, to exempt the testing industry from thorough independent review, regulation, and accountability.

3. A Vision of Testing as an Instrument to Enhance the Development of Human Talent

The fundamental recommendation of this Commission is that current testing policies and practices be substantially restructured to help people develop their talents and become constructive citizens and to help institutions become more productive, accountable, and just.

Specifically, we recommend that all uses of tests be evaluated with regard to how they contribute to or detract from the development of people as lifelong learners, and how they aid or hinder the work of institutions. Testing represents a social technology that, like most technologies, can be used properly or misused. Increased use of test results over the last several decades has resulted mostly from well-motivated efforts to improve schools, to make our social institutions more accountable, and to improve the capabilities of the nation's workforce. In striving to realize these goals, however, we have placed far more weight on educational and employment tests than any fallible technology can reasonably be expected to bear. A single test is too often asked to play many different roles. In some school systems, for example, the same test is expected to aid teachers' instructional decision-making, to diagnose students' learning problems, to place students in remedial programs, to determine grade promotion, to evaluate teachers, and to assess school quality. Likewise, one test is often used to assess persons of widely different backgrounds. For

example, a kindergarten readiness test is expected to evaluate equally well the potential of thousands of children with dissimilar cultural backgrounds, varied life circumstances, and diverse previous experiences with tests.

Although we cannot in this short report describe all the changes needed to achieve our vision of a more constructive role for assessment, we offer eight broad recommendations to guide the sorely needed restructuring of our nation's testing policies and practices.

1. Testing policies and practices must be reoriented to promote the development of all human talent.

In the past tests have been regarded mainly as selection devices or measuring instruments. We must move away from this view toward a vision of tests as instruments that can be used by individuals and institutions to promote human development. This shift requires that we re-evaluate the standards by which we judge the quality of test instruments, the names we give to tests, the ways we report test results, and, most of all, the ways we use such results.

First, to move beyond the view of tests as selection and classification instruments, the Commission recommends that we examine testing programs against the criterion of whether they are serving as gatekeepers or as gateways. No testing program should be tolerated if it leads to classification of people as "not able to

learn,” and shunts them into a dead-end situation. No testing program should continue if its benefits to an institution do not clearly outweigh its fiscal, social, and opportunity costs.

Second, if we take seriously the concept of tests as instruments to help people learn, then the names we attach to tests and the ways we report results should make learning more likely. Tests with names such as “general mental ability” or “scholastic aptitude” are much less useful in conveying what people can learn than are those with more accurate, specific, and understandable names such as “word knowledge” or “solving algebraic problems.”

Third, in practice, the principle that all test-based classifications should lead to opportunities also means that test results should be reported in terms of what examinees probably can and cannot do. Telling people *only* how their scores compare with the performance of some norm group will not suffice. Further, it is far more valuable to describe specific areas of strength and weakness in test performance than to provide one aggregate score representing performance over an uncertain mix of skills. No test score, or classification based on a test score, should be reported without some indication of

Assessment in the Service of Human Development

The notion that testing should be redirected to serve human development emerged quite early in the Commission’s work. It was argued perhaps most forcefully by Edmund Gordon in a keynote address at the hearing on testing and black Americans:

The critical problem for advanced societies is how to bring everyone...to a level of competence. It is not the question of how we select the people who can best fit into society. Because as I look down the road to the 21st century, I don’t think society will be able to afford a large segment of underdeveloped people....We are going to need all of our human resources, if not to support the economy, then to support democracy....The society that’s going to call itself democratic and is going to try to be democratic may have no choice but to invest its resources in this kind of universal development, and it may have to call upon its assessment industry to generate the kind of assessment data that informs resource development, not resource selection.

Gordon, E. (1988, December 2). Transcript of testimony presented at a hearing on Testing and the Allocation of Opportunities in Education and Employment for Black Americans, cosponsored by the National Commission on Testing and Public Policy. Washington, DC: Howard University.

the probable error associated with that score or classification. When volume makes it feasible test takers should also receive a detailed accounting of the questions they answered incorrectly. In line with our seventh recommendation below, such openness in reporting test results will help bring greater accountability to the enterprise of testing.

Finally, and most importantly, when people's opportunities and prospects are restricted because of classifications based in whole or in part on test results, programs should be available to help them overcome these restrictions. Thus, when a fifth grader is classified as "not ready" for the sixth grade, that classification should lead to appropriate learning opportunities. Similarly, when students are denied graduation from high school on the basis of a test — or for that matter on any other basis — they should have access to special tutoring or entry into an alternative high school education program, for example.

The Commission is *not* advocating stopgap remedies such as test coaching to get students over a testing hurdle, which is far too often

the present response to low test scores. The Commission emphatically is not recommending tracking in anticipation of test difficulties or as a result of students' scores on tests already shown to be fallible predictors. Instead we are calling for programs that go beyond elevating test scores to assist students in developing transferable academic and social skills, as assessed by multiple alternative indicators of learning or development.

This principle is reasonably easy to accept with regard to educational classifications of and decisions about young people. Since school attendance is mandatory, it seems only reasonable to require that a potentially harmful test-based classification and decision about a student be balanced by the availability of a demonstrably beneficial program. The Commission recommends, however, that this principle should extend analogously to all social institutions that use tests.

This means that potentially harmful test-based decisions about individuals made by government agencies or private-sector employers should be balanced by the availability of remediation efforts.

Commissioner Linn believes that because of the extraordinary range of types of tests and the uses of the test results, that it is worth noting that the applicability of Recommendation 1 to particular testing programs is quite variable. For example, the suggestion that test takers be given a detailed accounting of the questions answered incorrectly is reasonable in some contexts, but not others. Given currently available testing technology the relative benefits to test takers from such a requirement need to be weighed against the costs of generating sufficient alternate forms of the test and the feasibility of maintaining technical quality and equitable treatment of test takers across time

Central to our argument is the need to develop policies and deliver educational services that give the same importance to investing in our human resources that we give to investment in other forms of capital. The current national focus on economic competitiveness demands that we strengthen our collective efforts to maximize the potential of *all* the factors that contribute to our national economy — especially the talent of people. The current practice in accounting of recognizing only non-human capital investment should be changed so as to recognize investment in human capital through tax credits and similar institutional incentives. The Commission challenges organizations such as the Financial Accounting Standards Board to develop procedures for capturing the investment of employers in the continued development of their employees.

Employers who select and promote employees on the basis of a test should actively participate in educational collaboratives with public education, community-based organi-

zations, and other educational providers to close the employability gap for those who fail such workplace tests. Similarly, Congress should make available effective training programs designed to improve the knowledge and skills of young people deemed ineligible for military service on the basis of a selection test such as the ASVAB.

2. Testing programs should be redirected from overreliance on multiple-choice tests toward alternative forms of assessment.

For many purposes, direct assessment is preferable to indirect in any form. Just as a detective tries not to reach conclusions on the basis of only circumstantial evidence, important decisions about people and institutions ought not to be made on the basis of the circumstantial evidence of a single test. Instead, we should look to multiple sources of information, especially *direct* evidence of what we want the person to know or be able to do.

Toward Better Assessment in Employment

In a paper developed for the Commission, Richard R. Reilly and Michael A. Warech identify four alternatives for employee selection that “have been found to have equal or greater validity and less adverse impact than cognitive ability tests”

Trainability Tests: Presentation of standardized samples of job-related training materials followed by assesment of learning with paper-and-pencil or performance tests.

Work Samples: Performance of a task or set of tasks that have been shown (usually on the basis of a job analysis) to have direct, central relevance to the job in question. The tasks or job simulations are performed under standardized conditions.

Biodata: Biographical information collected in a standard application blank or a specially developed biodata form.

Assessment Centers: Use of a comprehensive set of assessment techniques according to standardized procedures, with emphasis on situational exercies and job-related simulation such as group problem-solving, fact-finding exercies, oral presentations, and role play.

Reilly, R.R. & Warech, M.A. (1988, December) *The validity and fairness of alternatives to cognitive tests*. Paper Prepared for the National Commission on Testing and Public Policy.

If we seek to develop specific skills or talents in our students, community members and employees, we ought to pay far more attention to actual demonstrations of such skills and talents than to performance on tests that are at best only indirect indicators of the talents of interest. In our schools, for example, we want students to learn how to write and to solve open-ended problems. Assessment should be based on actual samples of student writing rather than on multiple-choice tests in which students identify errors in writing passages. If we want students to learn how to solve open-ended science problems, we should assess their problem-solving skills by other means than multiple-choice tests in which they choose among alternative prescribed answers. Carefully crafted assessment devices would ask students to supply answers, perform observable acts, demonstrate skills, create products, and supply portfolios of work. If we want to certify persons who might be acceptable teachers, assessment should include their performance in classrooms as part of their teacher training or during a well-supervised internship period.

Alternative forms of assessment are desirable in general; and they are particularly essential in evaluating the competence of persons whose real-life skills may not be reflected in multiple-choice test results because their cultural or language backgrounds are different from the majority's. For these individuals, alternative forms of assessment would also help prevent unfairness. Fairness in the allocation of opportunities demands that important educational and employment decisions about people be based not on

one fallible indicator but on a range of relevant evidence.

The Commission finds no *one* existing form of testing or assessment to be universally better than group-administered multiple-choice tests. History shows that when other forms of testing — for example, written essays — were used by themselves to decide about people or institutions, they too eventually had negative effects on what students learned and what and how schools taught. Therefore, the Commission does not recommend that any one form of assessment replace current multiple-choice tests. Indeed, multiple-choice tests can provide us with useful information, and thus can be a legitimate part of many assessment programs. We reiterate, however, that no single form of assessment can shoulder the unbearable weight of being the sole measure of worth — or what passes for worth; and that, to enhance the educational and employment opportunities of minorities, various other forms of assessment *must* be included in any important decision-making about individuals, groups, and institutions.

3. Test scores should be used only when they differentiate on the basis of characteristics relevant to the opportunities being allocated.

Many of the most commonly used tests — namely, those that report a person's score by reference to the performance of persons in a comparison or "norm" group — are designed so that test scores differentiate among test takers. Sound test development should ensure that that differentiation is clearly related to the classifications and decisions being

made. If, for example, a test intended to measure students' math skills is shown to reflect how they do in English classes better than how they do in math classes, it is probably a poor test of math skills.

In employment testing, this principle has been recognized for more than two decades. Guidelines for employment testing require that tests used to allocate employment opportunities be shown to be related to the specific jobs for which people are being selected. Indeed, over the past twenty-five years, a number of uses of employment tests have been held to be illegal because they were *not* relevant to the jobs concerned.

Though the general principle of test relevance to the opportunities being allocated is widely accepted in the abstract, the standards for determining relevance are very often neglected. Far too often tests in both education and employment are accepted uncritically without an evaluation of whether they meet this criterion. At the most superficial level, for example, some test users do not look beyond the test title. They assume that a test called an "employment," "honesty," or "integrity" test will be relevant to the hiring of new employees. Educators may take for granted that a test of "critical thinking skills" is appropriate to select children for a gifted and talented program. But often a test's name does not accurately describe what it measures, much less indicate the meaning of a person's score, or the score's relevance to the opportunities being allocated. To assume that the test measures what its name suggests is itself uncritical thinking.

Even systematic examination of the content of a test may not be, by itself, an adequate basis for determining the relevance of test scores to the opportunities at issue. There are two reasons for this. First, the almost universal practice among test developers of having "validation" panels review the job relevance of the content of test items, while necessary, is not always sufficient. For example, many questions on teacher certification tests that survived such content scrutiny over the years have been roundly criticized by subsequent test reviewers as simplistic, containing professional shibboleths, having a shaky research base, or measuring trivial content. A test taker need not miss many such items before his or her classification, and hence eligibility to teach, is adversely affected.

Second and more broadly, as previously noted, test scores may be affected by a variety of circumstances and conditions not related to the test's content. For example, when a reading test is given under time constraints, scores may reflect test takers' ability to work quickly under pressure more than their reading skills *per se*. Thus, a person's score on a simple reading test — and hence inferences about how well the person reads — may change dramatically if the time allowed is increased from 30 minutes to an hour.

To ensure that test scores differentiate properly among test takers, evidence should be accumulated to show how well the test scores reflect real-life educational or job performance. While the exact bases for determining relevance to the opportunity concerned may vary with

test use, the Commission strongly recommends that such evidence be systematically accumulated and evaluated.

4. The more test scores disproportionately deny opportunities to minorities, the greater the need to show that the tests measure characteristics relevant to the opportunities being allocated.

Many testing programs adversely affect the educational and employment opportunities available to racial, ethnic, cultural, and language minorities and to women. As we have shown, this impact is due in part to broader social and educational inequalities that may be reflected in test scores and to biases in the tests. However, because tests are imperfect instruments, and because performance differences between minorities and non-minorities are often not as large as test score gaps, minorities are often disproportionately misclassified by tests. This runs directly contrary to our nation's commitment to provide equal educational and employment opportunities to all. Indeed, as the recent report of the National Research Council Committee on the General Aptitude Test Battery (GATB) pointed out, "fair test use would seem to require at the very least that the inadequacies of the technology should not fall more heavily on the social groups already burdened by the effects of past and present discrimination."

The vital importance of ensuring equality of educational and employment opportunities also requires that we adopt some immediate, but temporary, transitional strategies. For example, in reviewing the GATB the

National Research Council Committee recommended:

The ... use of score adjustments for black and Hispanic applicants in choosing which employment registrant to refer to an employer, because the effects of imperfect prediction fall more heavily on minority applicants as a group due to their lower mean test scores.

They went on to endorse:

The adoption of score adjustments that give approximately equal chances of referral to able minority applicants and able majority applicants; for example within-group percentile scores, performance-based scores or other adjustments.

The Commission recognizes that while helpful in the short run, the use of score adjustment strategies such as those recommended by the GATB Committee are not permanent solutions to the problem of ensuring equality of opportunity. It recognizes as well that the abandonment of a test-based allocation procedure will not automatically prevent its replacement by another selection scheme that is still less sound and equally discriminatory.

Thus, as a general long-term strategy we must as a society work toward using multiple forms of assessment, particularly direct as opposed to indirect, in allocating all important educational and employment opportunities. Further, purely statistical methods of detecting bias in our assessment instruments are altogether inadequate. Rather, it is essential to evaluate critically the fairness and accuracy of all test-based classifications in terms of the opportunities being allocated, with full awareness of the

implications for social groups already disadvantaged.

5. Test scores are imperfect measures and should not be used alone to make important decisions about individuals, groups, or institutions; in the allocation of opportunities, individuals' past performance and relevant experience must be considered.

On its surface, using a particular score on a test to allocate opportunities appears objective and fair: all who take the test must meet the same standard. But as we have shown, test scores used alone can result in wrong classifications about what people can or cannot do in the real worlds of school and the workplace. When such misclassifications trigger incorrect decisions, they create unfairness. Using tests as a tool for institutional accountability also seems sensible; but as we have seen, *overreliance* on tests distorts the information and diverts policy makers' attention from consideration of remedies for the underlying problems in the schools and workplace.

Thus, the Commission recommends that test scores *not* be used in isolation to make important decisions about people, groups, or institutions. Evidence that might contradict the conclusions based on a single test score must be considered. If a person who wants to dispute a speeding ticket based on a radar gun can offer evidence that the radar reading was inaccurate, or present other rebutting or mitigating evidence, test takers ought to have an analogous due process alternative.

While test scores may be useful for *informing* decisions about individuals, *past performance and experience relevant to the opportunities*

should be considered in conjunction with test scores. Traditions of fairness — as well as professional standards regarding test interpretation — demand that such decisions generally not be made on the basis of one kind of fallible evidence. Furthermore, *people and their judgments must be directly engaged* in making important decisions about other people. The Commission recognizes that human decisions about the allocation of opportunities can be biased and unfair, and that testing programs were often established to overcome such problems. Accumulated assessment evidence should be used, therefore, to evaluate the fairness and accuracy of human judgments in the allocation of opportunities.

6. More efficient and effective assessment strategies are needed to hold social institutions accountable.

A major cause for the distortion of test results and the ill effects of testing over the last several decades has been that the same test, or kind of test, has been asked to serve many important but different functions. Therefore, we recommend that testing for different purposes be differentiated and disentangled. Specifically, we urge that assessment of the effectiveness of social institutions — such as schools and training programs — be differentiated from assessment of individuals in order to help them.

Testing undertaken for purposes of institutional accountability calls for techniques different from those used to generate useful diagnostic feedback to teachers or students. While the latter need is not served particularly well by many present testing programs, we do have

useful techniques that can serve the former quite well. In the same way that public opinion polls sample public attitudes, testing for school or other institutional accountability should use a range of assessment devices across samples of people in those institutions. For example, the National Assessment of Educational Progress and several state assessments provide accurate measures of the learning of our nation's students by giving relatively small numbers of students a series of assessment exercises. Large school districts could benefit from similar sampling to gauge the performance of their schools. This strategy has the added bonus discussed in our second recommendation of using a much wider range of assessment instruments. It would also help prevent the distortions caused by using a single testing program to serve both instructional and accountability purposes, as noted above.

7. The enterprise of testing must be subjected to greater public accountability.

We have shown that although tests have become important instruments of public accountability, there are few mechanisms to *audit* or *appraise* the quality of publicly sponsored tests, to monitor their use as instruments of social policy, and to assess their impact on individuals, groups, and institutions. Enforcement of the Equal Employment Opportunity Commission guidelines is linked to expensive, time-consuming litigation. Standards for educational and psychological tests developed by professional organizations, while very useful, lack any effective enforcement mechanisms. Thus, the Commission recommends the development of additional institutional means to examine the quality of

tests and assessment instruments and to provide oversight of test use. Standards of accountability are most needed for tests supported by public funds and used to allocate opportunities.

While we cannot recommend one particular mechanism as the single best way to ensure accountability, several approaches are possible. First, some form of governmental scrutiny might be considered. In the same way that the federal Food and Drug Administration helps to protect the public against unsafe and ineffective drugs, a federal test bureau might help to protect against faulty tests or flawed uses of tests.

At the same time, we realize the limits of government regulation. The seriousness of federal scrutiny can waver with the political preferences of the administration in office. Further, some of the clearest misuses of testing — such as the use of tests in education to control school entry, grade promotion, and graduation or to determine teacher merit pay — have been a result of government action.

A second option, then, is some form of independent quality control, perhaps modeled on the practices of the Consumers' Union or the Underwriters Laboratory, to evaluate the technical quality of tests and the ways they are used.

If there is a specific perspective on testing that must be represented in such an independent agency, it is that of the groups who have been most adversely affected by testing in the past, namely, the ethnic, cultural, and language minorities of the nation. To realize this goal, the government and foundations might provide financial support and other incentives to help

minorities enter the field of testing. Training programs will need not only to develop a high degree of technical skills in such individuals, but also to concentrate on wide-ranging, non-technical aspects of the testing enterprise, including social, political, ethical, legal, and economic issues.

Finally, in the interests of promoting greater accountability in testing, and of using assessment to enhance the development of human talent, we recommend that far greater openness be brought to the enterprise. Thus, as noted in our initial recommendation, tests need to be more accurately labeled, results more constructively reported, and evidence as to what tests do and do not measure made more accessible.

8. Research and development programs must be expanded to create and use assessments that promote the development of the talents of all of our peoples.

Our final recommendation is for a serious and systematic research and development effort around issues of testing and assessment. We need to learn much more about the limits and the potential of both current tests and new forms of assessment to identify and develop skills and talents.

Given the fallibility of tests and other assessments as devices for measuring human performance, we need research to help make our assessments more accurate. But we also need to develop methods to communicate effectively the uncertainty inherent in *all* assessment results. We need answers to questions such as: How can the names attached to tests and the manner in which results are reported be changed to communicate

correctly to test takers and users the fallibility and limits of results? How can results be reported so that the self-image of individuals as learners will be enhanced rather than diminished? How can the technology of testing be changed so that test takers receive not just scores but an explanation of which questions they got right and wrong and what problems they solved well or poorly? How can we help users of test results understand the costs associated with overreliance on one fallible indicator? How do individuals intelligently and sensitively use multiple sources of relevant information in making classifications or decisions and avoid reducing the richness of the information to a statistical formula?

Regarding matters of unfairness in the allocation of opportunities, we need to develop tests that help to illuminate the special talents and skills of people from different ethnic, cultural, and language backgrounds. No test can be wholly free of cultural bias, for as products of culture, tests are permeated with cultural implications in both form and content. We must stop pretending that any single standard test can illuminate equally well the talents and help promote the learning of people from dramatically different backgrounds. Toward this end, individuals of different ethnic, cultural, and language backgrounds must be recruited not just to review existing tests, but to work in testing companies that build tests, to help assure that they are appropriate in form and content.

More broadly, we need to develop techniques to help people and society generally to understand and evaluate the kind and extent of errors that occur when people are ranked or classified. There never will — and

never should — be strictly technical answers to how different kinds of classification errors ought to be weighed. The balance of concerns over false rejections versus false acceptances is surely different when we are dealing with school children than when we are considering adults for occupational licensing. Unfortunately, techniques have not yet been developed to help decision makers and the public understand and weigh such tradeoffs.

Additionally, we need to develop standards and techniques to ensure that opportunities allocated through the use of tests — or any other indicators, for that matter — are not based, either explicitly or implicitly, more strongly on people's race, culture, national origin, language, or gender than on characteristics relevant to and rightly associated with those opportunities. In other words, we need to develop standards to prevent unfair racial, linguistic, and gender bias from creeping not just into

our test items, but, more importantly, into the classifications based on test scores and other data about people.

If we are to prevent tests from undermining social institutions and policies, we need to look far more closely at how tests and other forms of assessment affect people and institutions. We know that a test score used to label a child as mentally retarded communicates more about the limits than the potential of the child. How then can results of tests and other assessments be conceived and interpreted so that visions of human potential are enlarged rather than circumscribed?

In summary, we need to support research and development not just to refine old instruments or develop new ones. Rather, we need to find better ways to use assessment constructively and fairly to identify talent and to help develop people, assist institutions, and justly allocate opportunities.

A Federally Mandated R&D to Develop Culturally Specific Tests

The role of social, cultural, and linguistic factors in determining test performance is now recognized in federal law. Thus the section of Public Law 100-297 governing student assessment directs the Assistant Secretary for Indian Affairs to:

Establish and maintain a program of research and development to provide accurate and culturally specific assessment instruments to measure student performance in cooperation with the Tribes and Alaskan Native entities.

The principle of using culturally appropriate tests for assessment should be extended to other ethnic, linguistic, and cultural groups.

Source: P.L. 100-297, April 28, 1988, Sec. 5106.

4. A Final Word

In offering these recommendations, the Commission recognizes the importance, usefulness, and inevitability of testing in our society. As individuals and as a community, we will always need to know how our organizations and institutions are doing, what our children are learning and how well, and who among us are likely to make the most of opportunities that cannot be provided to all. If we are to answer these questions accurately and equitably, we must use all of the information that can be brought to bear.

The Commission believes that well-designed and responsibly used assessment instruments can be an important source of such information. If we succeed in redefining the mission of testing and directing it along a more constructive course, this information will be drawn from broader-based assessments that support rather than undermine individual and institutional goals, and that are appropriate to the purposes of the assessment. Assessments undertaken to inform instruction will be different from those designed to evaluate programs, to hold schools accountable for our children's learning, or to select good candidates for job training and advancement. The multiple-choice test will be de-emphasized. Moreover, we will expand the use of richer, more creative, and more varied devices that provide more direct evidence of the knowledge, skills, and behavior of interest in real-world settings.

Because well-designed instruments and the information gained from them can be used improperly, we must monitor and evaluate test use continuously. The Commission believes that assessment information should play an important but *suppor-*

tive role in decision-making, and that the critical factor in responsible practice is an active weighing and balancing of different forms of information. If the Commission recommendations to transform testing are heeded, assessment information will be used more soberly, more judiciously, and more responsibly. Except perhaps in matters of public health and safety, tests scores by themselves will never automatically trigger any action or decision, neither inclusion nor exclusion, neither reward nor sanction. Moreover, test-informed decisions will be differentiated clearly from absolute judgments: low test scores will never stigmatize an individual as a "failure" or permanently restrict the individual's life choices.

The Commission recognizes as well that meeting the challenge of the times will require more than a redirection of testing. New attitudes, policies, and practices related to testing can only support human development, not guarantee it. It will take resources as well as national resolve to realize these goals. While there is no expedient mechanism for altering the direction of social policies and practices, our history as a nation is filled with examples of how social institutions have been reshaped through a combination of thoughtful deliberation and bold effort on many fronts. The Commission calls on the nation, viewed around the world as "the land of opportunity," to mount yet another effort to bring testing policies and practices into line with our most important goals and deeply held convictions.

Reference Notes

THE COMMISSION'S REPORT HAS DRAWN broadly on Commission hearings and invited papers specially commissioned for our work, both described in the appendix, on other published research, and on personal and professional experiences of the members of the Commission. These notes do not attempt to document all of the material that has informed the work of the Commission, but instead simply provide explanation and documentation for specific facts and trends mentioned in our summary report.

Section 1. Introduction

Many observers over the last 30 years have recognized the transition of the American economy from one based primarily on land, labor, and capital as means of production to one based much more heavily on human services and information. Two widely noted authors on this subject have been Daniel Bell in *The Coming of Post-Industrial Society* (1976) and John Naisbett in *Megatrenda* (1982). The implications of this transformation for the increased importance of education have also been widely noted and have been given unprecedented national prominence by the effort of the National Governor's Association (NGA) and President Bush to establish national education goals. This effort began with an "education summit" in Charlottesville, Virginia, in the fall of 1989. In his state of the Union message in 1990, President Bush announced six broad goals for education, pertaining to students' readiness for school, student achievement in grades 4, 8, and 12, high school graduation rates, achievement in math and science, adult literacy, and freedom from drugs and violence in schools.

The NGA has continued to refine these goals, and a final version of has not yet been announced. Periodicals such as *Education Week* and the *Chronicle of Higher Education* have been covering the NGA's elaboration of the goals; see for example Walker (1990) and DeLoughry (1990).

Two of the most widely publicized recent studies of the education and training implications of demographic trends for the labor force in the United States are the Department of Labor's *Workforce 2000* report (1987) and *The Learning Enterprise*, jointly sponsored by the Department of Labor and the American Society for Training and Development (Carnevale & Gainer, 1989).

Projections of demographic trends appear regularly in the U.S. Bureau of the Census' Current Population Reports. According to one recent report in the series:

The share of the U.S. population under age 35 may never again be as large as it is now 55 percent. That percentage is projected to drop to 48 in 2000, 46 in 2010, and 41 in 2030.

The Black-and-other races population will experience a more substantial rate of growth. The Black population might increase by 14 million persons or 50 percent by 2030, while other races populations (primarily Asians, Pacific Islanders and American Indians) is projected to triple by 2040, growing by 16 million. (Spencer, 1989)

Exactly how much testing goes on in the United States is impossible to gauge precisely because so much testing is done by so many different agencies for so many different

purposes. Nevertheless, a variety of evidence indicates that reliance on tests has increased substantially over the last several decades, particularly in education. We estimate that the volume of testing has grown roughly 10 percent per annum from the mid-1950's into the 1980's, and some evidence suggests that the rate of increase has accelerated in the 1980's. While educational testing appears to have increased fairly steadily since the 1960's, the growth of employment testing appears to have slowed in the 1970's as a result of litigation. According to some sources it accelerated once again in the 1980's (Lee, 1988). Documentation on these trends, based on a variety of direct and indirect evidence, is provided in Haney, Madaus, & Lyons (in press). On the history of the AFQT see Sticht's (1989) invited paper, and on the history of U.S. Employment Service testing see Hartigan & Wigdor (1989). Other fairly general accounts of the history of testing are Resnick (1982), Hale (1982), Haney (1984), and Madaus & Kellaghan (in press); Ruger (1975), Nairn (1980) and Valentine (1987) provide organization-specific histories.

For elaboration on what is meant by a "test" and on the kinds of tests actually used in different realms of education and employment, see the forthcoming staff report of the Commission and volumes of research reports prepared for the Commission, published by Kluwer.

Section 2. Why Testing Must Be Transformed

The story of the Antonia Gonzalez case was provided by Albert Kauffman, consultant to the Commission. An-

tonia Gonzalez was not the woman's real name; we have changed it protect her confidentiality. However, all facts of this story are documented in a federal court case.

Finding 1: Tests are imperfect

Regarding the names given to tests, the 1966 *Standards for Educational and Psychological Tests* (American Educational Research Association, & National Council on Measurement in Education, 1966) warned that "Names given to published tests . . . should be chosen to minimize risk of possible misinterpretation by test purchasers and subjects" (p. 13). A major problem in this regard is that some of the terms used to name tests, such as intelligence, ability, and aptitude, have a narrow technical meaning in the field of testing that differs from usage in common parlance. Because of common misinterpretation of intelligence test scores, some tests previously labeled as "intelligence" tests no longer use this word in their names. For a discussion of possible misinterpretation of the word "aptitude" in the Scholastic Aptitude Test (SAT), see Crouse & Trusheim (1988).

Regarding the predictive power of tests, in the text of its report the Commission has avoided technical discussions of prediction equations, variance explained, and adjustments of parameters of prediction equations for factors such as restriction of range and unreliability in test scores or criterion measures. For a theoretical discussion of such measurement issues, see Lord and Novick (1968). For a discussion in the context predictive validity of tests, see Linn (1982); and for a more recent discussion of the predictive validity of employment tests see Hartigan &

Wigdor (1989). Linn (1982, p. 339) points out that many studies of the power of tests to predict teacher's grades have shown that the correlations between test scores and test grades average around 0.50. This means that the amount of variance in the criterion of school grades "explained" by tests scores is around 25 percent (since the square of the correlation coefficient yields the amount of variance explained). As Linn points out, "Even with a correlation as high .60, some persons with quite low test scores can be expected to rank among the top fifth on the criterion. Conversely, the expectation is that a few people with very high test scores will not rank even among the top 80 percent on the criterion" (p. 337). A large number of studies of the ability of college admissions tests to predict college freshman grades has shown that the correlation between admissions tests and first year college grades tends to be about 0.35, meaning that admissions tests explain about 12 percent of the variance in freshman grades. Research has generally shown that high school grades can predict college grades as well as or better than college admissions tests (see Linn, 1982; Donlon, 1984). When both high school grades and tests are used together, the pair can predict college freshman grades better than either one used alone. In contrast to this general literature on predictive power college admissions tests, Crouse & Trusheim (1988) have analyzed the value of admissions tests to help make categorical admissions decisions (admit or not admit) and have concluded that admissions test scores do not add greatly to improving the accuracy of such classifications in general and their use in contrast to the

use in high school grade alone tends to have adverse impact on both low-income and black applicants.

The recent report of National Research Council (Hartigan & Wigdor, 1988) on the U.S. Employment Services' General Aptitude Test Battery (GATB) recounts that the validity coefficient of the GATB for predicting training criteria (uncorrected for possible unreliability in the criteria of training) averages around 0.35 overall and 0.30 for studies conducted since 1972. In comparison, the mean uncorrected validities of the Armed Forces Qualifying Test for predicting training criteria was estimated to average 0.33 across all services (Hartigan & Wigdor, 1989, pp. 88-98). The validity of employment tests for predicting on-the-job performance, typically in the range of 0.10 to 0.30, has generally been found to be less than their validity for predicting training criteria. In part, this finding results from the unreliability in criterion measures of job performance, such as supervisor's ratings. Nevertheless, the general finding that tests predict less than about one-quarter of the variance in actual on-the-job or in-school performance is clear from research literature dating back to the 1920's (a correlation of 0.50 between a test and a criterion is explained by the test).

There is a vast literature on the influences of environmental conditions on test performance. Standard textbooks on testing (such as Cronbach, 1984, and Anastasi, 1988) discuss some of the older pertinent literature. For several studies on how cognitive and linguistic factors may influence test performance, see *Cognitive and Linguistic Analyses of Test Performance* (Freedle & Duran, 1987). For a sum-

mary of the literature on test coaching and preparation see Kulik, Bangert-Drowns, & Kulik (1984) and section 6 of Haney, Madaus, & Lyons (in press). For a discussion of how such environmental conditions may influence the test performance of minority individuals see Samuda (1975). For specific studies on the effects of eating breakfast and of noise on test results, see Meyers, Sampson, Weitzman, Rogers, & Kayne (1989), Bronzaft & McCarthy (1975), and Bronzaft (1981, 1985).

The example of classifying 1000 people on the basis of job performance and test scores assumes a correlation between test scores and performance of 0.35, and the results mentioned in the text are derived from a Monte Carlo simulation of the situation describes. For a theoretical discussion of this classification problem see Lord & Novick (1968, pp. 275-277, 344-347).

For discussion of the problem of setting cutoff scores on tests see Glass (1978), Hambleton (1978), and Madaus (1985b). While some decision-theoretic approaches for weighting different kinds of classification errors have been propose in theory (Cronbach & Gleser, 1965; Petersen & Novick, 1976), we know of no instances in which such models have actually been used in making decisions about classification based on test scores.

For a description of the ASVAB misnorming, and a summary of studies performed to assess the performance of recruits "mistakenly" admitted as a result of the misnorming see Sticht (1989, invited paper). Sticht's summary of the studies reviewed is:

Overall, the PI's [that is the recruits "mistakenly" admitted

because of the miscalibration] performed slightly less well than the controls. Yet the difference were not large, and in several cases the PIs performed as well or better than the controls of average aptitude. Performance differed according to military service studied. Whereas in the Army, there was often not much difference . . . in the Air Force and Navy the differences among the PIs and higher aptitude groups were larger. This presumably reflects the differences in personnel policies such as the requirements in the Air Force and Navy to study job manuals and pass paper-and-pencil tests for promotion to higher paygrades. In general, whenever lower aptitude personnel are place in a test-taking situation, they tend to do more poorly than when their work is evaluated by supervisor ratings. (p. 61)

Finding 2: Some test uses result in unfair treatment. . .

On the accuracy of classifications based on readiness tests see Shepard & Smith (1988b), Cunningham (1989), and the invited paper that Cunningham prepared for the Commission.

The Commission recognizes that issues of test fairness and bias have been treated extensively in the research literature. In much of this literature these issues have been addressed primarily in terms of the regression relationships between test scores and various criteria of performance (such as grade averages for educational tests and measures of training or employment success for employment tests). This approach to

studying test bias is sometimes called the Cleary model after Cleary (1968). In general such analyses have not found statistically significant differences for minority and non-minority individuals. See Linn (1982) for a summary of some of this literature.

The Commission has not adopted the regression analysis perspective on test bias and fairness because it has been primarily interested in the use and misuse of test data in the allocation of opportunities that is, in the use of test scores not just making slightly differentiated predictions about individuals, but in making categorical decisions or classifications as to whether individuals are accepted or rejected for opportunities (e.g., admission to education programs or schools, or acceptance into the military or entry-level employment).

The fairness of using test results to make accept/reject decisions has been examined in the research literature primarily in terms of what have come to be called group parity models of fair selection. These models have sought to address the issue of how cutoff scores may be set on tests so that fairness may be achieved between two groups that score differently on the test. In one model, Thorndike proposed that "the qualifying scores on a test should be set at levels that will qualify applicants in the two groups in proportion to the fraction of the two groups reaching a specified level of criterion performance" (Thorndike, 1971, p. 63). In subsequent literature this group parity model of fair selection came to be called the constant ratio model, because it suggests that cutoff scores should be set so that for different groups there is an equal or constant ratio between the numbers of

persons selected on the basis of the test and numbers likely to be successful on the criterion. Alternatively, Cole (1973) proposed what is now called the conditional probability model of fair selection tests (or other selection devices) are set so that for individuals in different groups, for those likely to succeed there should be equal probability of selection regardless of group membership. Linn (1973) proposed a third group parity model, which came to be called the equal probability model. It suggested that cutoff scores for different groups should be set so that for those in different groups who fall should be equal proportions likely to succeed, regardless of group membership. Though each of these models of fair selection has intuitive appeal, commenters have noted that they (and their logical converses) actually imply different cutoff scores for particular sets of data on test and criterion performance (see for example, Petersen & Novick, 1976, and Jensen, 1980). For a practical example of how these different perspectives can yield very different pictures on the matter of fairness in testing and selection see pp. 258-260 of Hartigan & Wigdor (1989).

The perspective adopted in the text of the Commission report is analogous to Cole's conditional probability model. The reason for this perspective on fair selection is our concern that more weight needs to be placed on direct indicators of performance in school or on the job than on test performance. Cole's model, in contrast to the equal probability model, employs as a denominator the number or proportion of those likely to succeed on the performance criterion. Nevertheless, the Commission wants to make clear that it is not endorsing the

conditional probability model as a prescription for ensuring fairness in selection. Fairness is ultimately an ethical matter that is not reducible to a statistical model. And more specifically the Commission has argued against the use of any fixed cutoff scores on tests in making decisions about individuals, regardless of whether or not different cutoff scores are set for different groups.

Sources showing the general pattern of lower test scores for black Americans, Hispanics, American Indians, and women include College Board (1988), Graduate Management Admissions Council (1988), Hartigan & Wigdor (1989), and Jaynes & Williams (1989). See also the invited papers prepared for the Commission by Chavers and Locke, Kochman, Moore, Tsang, Valdes, Duran, Jones, Pennock-Roman, Rosser, and a hearing report by Salinas Sosa. Specific sources of the data presented regarding differential test performance are Dean Chavers' testimony at the Commission hearing on the effects of testing on American Indians; College Board (1988), American College Testing Program (1988), Florida State Department of Education (1989), Graduate Management Admissions Council (1988), and testimony from the Commission hearing regarding Asian and Pacific Americans listed in the Appendix. Regarding passage rates on so-called teacher competency tests see also Jaynes & Williams (1989, p. 364) and Haney, Madaus, & Kreitzer (1987).

The quotation from the National Research Council report comes from Jaynes & Williams (1989, p. 346). Regarding group differences on indicators of performance other than tests see College Board (1988), Crouse & Trusheim (1988), and the

invited paper prepared for the Commission by Reilly & Warech. The National Research Council report example concerning classification of carpenters in Hartigan & Wigdor (1989, 258-260).

Regarding alternative forms of assessment see Crouse & Trusheim (1988) who show that high school grades have equal validity for making college admissions decisions with less adverse impact on black American and lower income applicants than the SAT college admissions test. Regarding employment tests see the invited paper by Reilly & Warech listed in the Appendix and also Reilly & Chao (1982). For other sources on alternative assessment see the invited papers by Haertel; Brown; Campione, Webber & McGilley; Resnick & Resnick; Gardner; and Sternberg, all listed in the Appendix. Though minority groups and women in general would be less adversely affected if opportunities were allocated via some kinds of alternative assessments, it should not be assumed that all alternatives to standardized multiple-choice tests would have this effect. For example, data in College Board (1988) show that while women, American Indians, black Americans and Mexican Americans would gain if college admissions decisions were based on the Test of Standard Written English rather than SAT scores, this is not the case for Asian Americans and Puerto Rican Americans.

Finding 3: Students are subjected to too much testing . . .

Many of the specific facts and estimates reported here are drawn from Haney, Madaus, & Lyons (in press). The *Education index data* were previously presented in (Haney, 1986), but are updated in Haney, Madaus,

and Lyons (in press). The data on state testing programs are drawn from the Office of Technology Assessment (December, 1987). The data on the reported dollar volume of tests sales are drawn from the Bowker report series (R.R. Bowker Company, 1970-1987), but have been shown to be incomplete in Haney, Madaus, & Lyons (in press). The other specific studies referred to regarding this finding are Romberg, Zarinnia, & Williams (1989), Cunningham (1989), Haladyna, Nolen, & Hass (1989), Koretz (1988), Madaus (1985b), and Madaus & Kellaghan (in press). See also Medina & Neill (1988).

Finding 4: . . . undermining important social policies . . .

Specific literature directly referred to in connection with this finding includes Cannell (1987, 1988, 1989), Linn, Graue, & Sanders (1989, 1990), Shepard (1989), and Hartigan & Wigdor (1989). Regarding the trickle-down effect of tests on early child educational programs see Cunningham (1988) and Shepard & Smith (1988a). The American Business Conference proposal for a new test for all high school graduates was reported in the *Washington Post* (1989). The example regarding the New Jersey Job Services Office comes from Hartigan & Wigdor (1989, p. 216).

Finding 5: Public Accountability

Over the last 30 years there have been numerous calls for greater accountability in the testing industry. Some of the most prominent relevant works are Hoffman (1962), Houts (1971), Nairn (1980), Owen (1985), Cannell (1987), and Cannell (1989). Though the Commission has not endorsed a particular means of ensuring greater public accountability with regard to testing, we note with

interest that the executive director of the Commission, George Madaus, is currently undertaking a study of alternative models by which more active monitoring by publicly sponsored testing programs might be achieved. For background on the motivation for this study and the need for greater accountability regarding testing, see Madaus (1985a).

On the history of federal guidelines and professional standards for testing see Novick (1982), Haney & Madaus (in Press), Hartigan & Wigdor (1989), and invited papers listed in the Appendix by Chachkin, Gelerter, Patterson, Rebell, Clune, McDowell & Dodge, and Wiesen, Abrams & McAtree.

Section 3. A Vision of Testing as an Instrument to Enhance the Development of Human Talent

The Commission's first recommendation is essentially a summary of its vision of how testing must be transformed.

Regarding the second recommendation see invited papers listed in the Appendix by Haertel; Brown; Campione; Webber & McGilley; Resnick & Resnick; Reilly & Warech; Gardner; and Sternberg. For older literature that argues for less reliance on multiple-choice tests and more reliance on direct forms of assessment see Hoffman (1962), Houts (1971), and Madaus & Kellaghan (in press).

The Commission's recommendations 3 and 4 essentially call for evidence of the validity of classifications based on test results. What we are advocating is referred to in the research literature as more predictive (as opposed to mere content) validity evidence, and evidence that is directed not just at the validity of test scores in the abstract but more specifically at the

classifications and opportunity allocations informed by test scores. Moreover, we argue that the greater is the adverse impact of testing programs on minorities, the greater is the need to provide not just content but also predictive validity evidence. The quotation from the NRC review of the GATB comes from Hartigan & Wigdor (1989 p. 260).

Recommendations 5 and 6 are clearly supported by American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1985 *Test Standards*), and by evidence cited above showing that some current uses of the test results, in isolation from other sources of evidence about student learning and educational programs, are having the effect of undermining important social policies. Specific quotations from the NRC Committee on the General Aptitude Test Battery come from Hartigan & Wigdor (1989, pp. 11-12).

The Commission's call for additional research and development on assessment (recommendation 8) is somewhat similar to previous calls for research and development on testing and assessment, such as those made in Tyler & White (1979), but with two major differences. First, the Commission argues that the touchstone in new development efforts ought to be how assessments aid in the development and utilization of human talent. Second we direct special attention not just to the technical aspects of new tests and assessments but more importantly to how people and institutions use and interpret results from such assessments and how such assessments affect people's opportunities particularly the opportunities of groups that have not had

access to equal educational and employment in the past.

Bibliography

This bibliography does not include all of proceedings of Commission hearings or invited papers listed in the Appendix, except for some that are already published or in press.

- American College Testing Program. (1988). *Executive summary, National ACT Assessment results 1987-88*. Iowa City, IA: American college Testing Program.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1985). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.
- American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1966). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.
- Anastasi, A. (1988). *Psychological Testing* (6th ed.). New York: Macmillan.
- Bell, D. (1976). *The coming of post-industrial society*. New York: Basic Books.
- Bronzaft, A. (1981). The effect of a noise abatement program on reading ability. *Journal of Environmental Psychology*, 1, 215-222.
- Bronzaft, A. (1985). Combating the unsilent enemy—noise. *Prevention in Human Services*, 41 (1-2), 179-192.
- Bronzaft, A., & McCarthy, D. (1975). The effect of elevated train noise on reading ability. *Environment and Behavior*, 7(4), 517-527.
- Cannell, J.J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average*. Albuquerque, NM: Friends for Education.
- Cannell, J.J. (1988). *Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average*. *Educational Measurement: Issues and Practice*, 7(2), 5-9.
- Cannell, J.J. (1989). *The 'Lake Woebegon' report: How public educators cheat on standardized achievement tests*. Albuquerque, NM: Friends for Education.
- Carnevale, A., & Gainer, L. (1989). *The learning enterprise*. Alexandria, VA: American Society for Training and Development Press.
- Cleary, T.A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement*, 5, 115-124.

- Cole, N. (1973). **Bias in selection.** *Journal of Educational Measurement*, 10, 237-255.
- College Board. (1988). *National college-bound seniors: 1988 profile. Profiles of SAT and Achievement Test takers. National Ethnic/Sex profile.* New York: The College Board.
- Cronbach, L.J. (1984) *Essentials of psychological testing (4th ed.)* New York: Harper & Row.
- Cronbach, L. J. & Gleser, G.C. (1965). *Psychological tests and personnel decisions (2d ed.)*. Urbana: University of Illinois Press.
- Crouse, J., & Trusheim, D. (1988). *The case against the SAT.* Chicago: University of Chicago Press.
- Cunningham, A.E. (1989). *Testing in early childhood. The Educator (January), 8-9.*
- DeLoughry, T. (1990, March 7). **Governors approve education goals with little mention of colleges.** *Chronicle of Higher Education*, p. A24.
- Donlon, T. (Ed). (1984). *The College Board handbook for the Scholastic Aptitude test and Achievement Tests.* New York: College Entrance Examination Board.
- Florida, State Department of Education (1989). *March 1989 CLAST Results.* Tallahassee. FL.
- Freedle, R., & Duran, R. (1987). *Cognitive and linguistic analyses of test performance.* Norwood, NJ. Abex.
- Glass, G.V. (1978). **Standards and criteria.** *Journal of Educational Measurement*, 15, 237-261.
- Graduate Management Admissions Council. (1988). *An admissions office profile of candidates taking the Graduate Management Admission Test 1983-84 through 1987-88.* Princeton, NJ: Graduate Management Admissions Test Program Direction Office.
- Haladyna, T.M., Nolen, S.B., & Hass, N. S. (1989). *Report to the Arizona State Legislature: Test score pollution.* Phoenix, AZ: Arizona State University West Campus.
- Hale, M. (1982). *History of employment testing.* In a. Wigdor & W. Garner (Eds.), *Ability testing: Uses, consequences and controversies, Part II (pp. 3-38).* Washington DC: National Academy Press.
- Hambleton, R.K. (1978). *On the use of cutoff scores with criterion referenced test in instructional settings.* *Journal of Educational Measurement*, 15, 277-290.
- Haney, W.M. (1984). *Testing and reasoning and reasoning about testing.* *Review of Educational Research*, 54(4), 597-654.

- Haney, W.M. (1986). College admissions testing and high school curriculum: Uncertain connections and future directions. In *Measures in the college admissions process: A College Board colloquium* (pp. 32-51). New York: College Board.
- Haney, W.M., & Madaus, G.F. (In press). The evolution of ethical and technical standards for testing. In R. Hambleton (Ed.), *Handbook of testing*. Amsterdam: North-Holland Publishing Company.
- Haney, W.M., & Madaus, G.F. & Kreitzer, A. (1987). Charms talismanic: Testing teachers for the improvement of American education. *Review of Research in Education*, 14, 169-238.
- Haney, W.M., & Madaus, G.F. & Lyons, R. (In press). *The fractured marketplace for standardizing testing*. Boston: Kluwer.
- Hartigan, J. Widgor, A.K. (1989). *Fairness in employment testing: Validity generalization, minority issues and the General Aptitude Test Battery*. Washington, DC: National Academy Press.
- Hoffmann, B. (1962). *The tyranny of testing*. New York: Crowell-Collier.
- Houts, P. (1971). *The myth of measurability*. New York: Hart.
- Jaynes, G.D., & Williams, R.M. (1989). *A common destiny: Blacks and American society*. Washington, DC: National Academy Press.
- Jensen, A.R. (1980). *Bias in mental testing*. New York: Free Press.
- Koretz, D. (1988). Arriving in Lake Wobegon: Are standardized test exaggerating achievement and distorting instruction? *American Educator*, 12(2), 8-15, 46-52.
- Kulik, J.A., Bangert-Downs, R.L., & Kulik, C.-L.C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin*, 95(2), 179-188.
- Lee, C. (1988). Testing makes a comeback. *Training*, 25(2) 49-59.
- Linn, R.L. (1973). Fair test use in selection. *Review of Educational Research*, 43, 139-161.
- Linn, R.L. (1982). Ability testing: Individual differences, prediction and differential prediction. In A.K. Wigdor, & W.R. Garner (Eds.), *Ability Testing: Uses, consequences and controversies, Part II* (pp. 335-388). Washington, DC: National Academy Press.
- Linn, R.L., Graue, M.E. and Sanders, N.M. (1989 March). Comparing state and district test results to national norms: Interpretations of scoring 'above the national average'. Paper presented at the annual meeting of the

- American Educational Research Association, San Francisco, CA.
- Linn, R.L., Graue, E.M., & Sanders, N.M. (1990). *Comparing state and district test results to national norms: Interpretations of scoring 'above the national average'*. (CSE Technical Report 308). Los Angeles: University of California at Los Angeles, Center for Research on Evaluation, Standards and Student Testing.
- Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Madaus, G. (1985a). Public policy and the testing profession—You've never had it so good. *Educational Measurement: Issues and Practice*, 4(4), 5-11.
- Madaus, G. (1985b). Test scores as administrative mechanisms in educational policy. *Phi Delta Kappan*, 66(9), 611-617.
- Madaus, G. & Kellaghan, T. (in press). Curriculum, evaluation and assessment. In P.W. Jackson (Ed.), *Handbook of Research on Curriculum*. Washington, DC: American Educational Research Association.
- Medina, N., & Neill, D.M. (1988). *Fallout from the testing explosion*. Cambridge, MA: FairTest.
- Meyers, A., Sampson, A., Weitzman, M., Rogers, B., & Kayne, H. (1989). School breakfast programs and school performance. *American Journal of Diseases of Children*, 143, 1234-1239.
- Nairn, A. (1980). *The reign of ETS: The corporation that makes up minds: The Ralph Nader Report*. Washington, DC: Learning Research Project.
- Naisbett, J. (1982). *Megatrends: Ten new directions transforming society*. New York: Warner Books.
- Novick, M. (1982). Federal guidelines and professional standards. In A.K. Wigdor & W.R. Garner (Eds.), *Ability testing: Uses consequences and controversies, Part II* (pp.70-98). Washington, DC: Office of Technology Assessment, U.S. Congress.
- Owen, D. (1985). *None of the Above: Behind the myth of the SAT*. New York: Houghton Mifflin.
- Petersen, N.S., & Novick, M.R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, 13, 3-29.
- R.R. Bowker Company. (1970-1987). *The Bowker annual of library and book trade information*. New York: Author.
- Reilly, R.R., & Chao, G.T. (1982). Validity and fairness of some alternate employee selection procedures. *Personnel Psychology*, 35, 1-62.

- Resnick, D. (1982). History of educational testing. In A. Wigdor & W. Garner (Eds.), *Ability testing: Uses, consequences and controversies, Part II* (pp. 173-194). Washington, DC: National Academy Press.
- Romberg, T.A., Zarinnia, E.A., & Williams, S.R. (1989). *The influence of mandated testing on mathematics instruction: Grade 8 teachers' perceptions*. Madison: University of Wisconsin, National Center for Research in Mathematical Science Education.
- Ruger, M.C. (1975). *A history of the American College Testing Program (1959-1974)*. Doctoral dissertation. University of North Colorado.
- Shepard, L. (1989, March). *Inflated test score gains: Is it old norms or teaching to the test?* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Shepard, L., & Smith, M.L. (1988a). Escalating academic demand in kindergarten: Counterproductive policies. *Elementary School Journal*, 89(2), 135-145.
- Shepard, L., & Smith, M.L. (1988b). Flunking kindergarten: Escalating curriculum leaves many behind. *American Educator*, 12(2), 34-39.
- Spencer, G. (1989). *Projections of the population of the United States, by age, sex and race: 1988-2080*. (Current Population Reports, Series P-25, No. 1018). Washington, DC: U.S. Government Printing Office.
- Standardized test planned to determine work skills. (1989, November 6). *Washington Post*.
- Thorndike, R.L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement*, 8, 63-70.
- Tyler, R., & White, S. (1979). *Testing, teaching and learning: Report of a conference on research on testing*. Washington, DC: National Institute of Education.
- U.S. Department of Labor. (1987). *Workforce 2000*. Washington, DC: U.S. Government Printing Office.
- Valentine, J.A. (1987). *The College Board and the school curriculum*. New York: College Entrance Examination Board.
- Walker, R. (1990, February 28). Governors set to adopt national education goals. *Education Week*, p. 16.

Appendix:

National Hearings *

THE NATIONAL COMMISSION ON TESTING and Public Policy is indebted to the following individuals and organizations for planning and hosting the five national hearings that examined the impact of testing on minorities and women: the Native American Scholarship Fund, Inc. Dr. Dean Chavers, Dr. Gladys Levis-Pilz, Dr. Tony Lam, and Ms. Sharon Wilson of the University of New Mexico, for the hearing on American Indians; the National Association for Asian and Pacific American Education, Dr. Amy Agbayani, and Dr. Morris K. Lai of the University of New Mexico, for the hearing on Asian and Pacific Americans; Dr. Sylvia Johnson, Dr. Wilfred Jonson, Ms. Christine Phillips, and Ms. Tina Farrell of the Department of Psychoeducational Studies, Howard University, for the hearing on Black Americans; Dr. José Cárdenas, Dr. Gloria Zamora, and Dr. Alicia Salinas Sosa of Intercultural Development Research Association for the hearing on Hispanics; and Ms. Marcia D. Greenberger, Ms. Ellen Vargas, Ms. Katherine Conner, and Ms. Laura Epstein of the National Women's Law Center for the hearing on gender bias.

The Impact of Testing on Hispanics

Co-sponsored by Intercultural Development Research Association (IDRA)
San Antonio, Texas
February 26-27, 1988

Invited Speakers

Richard P. Duran, University of California at Santa Barbara, Santa Barbara, California: Testing of Hispanic students: Implications for secondary education

Respondent:

Richard A. Figueroa, University of California at Davis, Davis, California

Maria Pennock-Roman, Educational Testing Service, Princeton, New Jersey: The status of research on the Scholastic Aptitude Test (SAT) and Hispanic students in postsecondary education

Respondents:

Charlene Rivera, Development Association, Arlington, Virginia
John Weiss, National Center for Fair and Open Testing, Cambridge, Massachusetts

Testimony

Teacher Education

Cynthia Darche Park, San Diego State University, San Diego, California
Peter A. Garcia, Saint Mary's College of California, Moraga, California

* Proceedings of the five hearings will be made available through ERIC Clearinghouse on Urban Education (CUE), Teachers College, Columbia University

Graduate Admissions

Judith Walker de Felix, University of Houston, Houston, Texas
W. Robert Houston, University of Houston, Houston, Texas
Robert D. Milk, The University of Texas at San Antonio, San Antonio, Texas

Other Higher Education

James A. Vazquez, University of Washington, Seattle, Washington
Nancy Wainstein, University of Washington, Seattle, Washington
Sally Mitchell, Incarnate Word College, San Antonio, Texas
Richard J. Harris, The University of Texas at San Antonio, San Antonio, Texas

Employment

Richard S. Barrett, Barrett Associates, Hastings on the Hudson, New York

Testing, Opportunity Allocation, and Asian and Pacific Americans

Co-sponsored by the National Association for Asian and Pacific American Education
Honolulu, Hawaii
April 11, 1987

Invited Speakers

Chui Lim Tsang, Chinatown Resources Development Center, San Francisco, California: Cultural and linguistic sources of bias and unfairness in the informal assessment of Asian Americans

Testimony

Sheila Forman, Hawaii Mental Health Association, Honolulu,

Hawaii: The testing of prospective teachers in Hawaii

Selvin Chin-Chance, Hawaii State Department of Education, Honolulu, Hawaii: Test scores as indicators of educational problems

Ormond W. Hammond, Office of Program and Planning, Kamehameha Schools, Honolulu, Hawaii: The assessment of the education of native Hawaiians

Tony C. M. Lam, Department of Educational Foundations, University of New Mexico, Albuquerque, New Mexico: Modification of standardized tests administered to special Asian and Pacific American populations

Pio DeCano, Tacoma Public Schools, Tacoma, Washington: Testing and the identification of appropriate educational services for limited-English-proficient Asian and Pacific American students

Rona Rodenhurst, Hawaii State Office of Hawaiian Affairs, Honolulu, Hawaii: Testing and the educational status of native Hawaiians

Rosita Galang, National Association for Asian and Pacific American Education, San Francisco, California: Research development, and, policy issues pertaining to tests, testing, and Asian and Pacific Americans

Edmund Lee, California Association for Asian and Pacific Bilingual Education, Los Angeles, California: Testing and Asian and Pacific Americans in California

John Lum, San Francisco Unified School District, San Francisco, California: Testing research and test taker rights

Robert Hall, Hawaii Institute for Biosocial Research, Honolulu, Hawaii: Ethics of Educational testing

Thomas T. Saka, University of Hawaii at Manoa, Honolulu, Hawaii: Test scores and the Ethnic Minority Index

Testing and the Allocation of Opportunities in Education and Employment for Black Americans

Co-sponsored by the Department of Psychosocial Studies, School of Education, Howard University Washington, DC

December 2-3, 1988

W. Curtis Banks, Howard University, Washington, DC: Methodological criticisms in fair employment testing: Black deconstruction in America

Respondents

Barbara Williams, Washington, DC

Ayers D'Costa, Ohio State University, Columbus, Ohio

Edmund Gordon, Yale University, New Haven, Connecticut: The changing purposes of psychometrics in human resource development

Respondents

William Hall, University of Maryland, College Park, Maryland

Stephanie Shipman, U.S. General Accounting Office, Washington, DC

Valarie Ford, District of Columbia Public Schools, Washington, DC: Test-based decisions in secondary schools and beyond

Respondent

Charles Asbury, School of Education, Howard University, Washington, DC

Linda Darling-Hammond, The RAND Corporation, Washington, DC

Lloyd Bond, University of North Carolina at Greensboro, Greensboro, North Carolina: Understanding the Black and White student gap on measures of quantitative reasoning

Sukai, Prom-Jackson, Washington, DC

Leroy Tompkins, Office of Educational Accountability, Montgomery County Public Schools, Rockville, Maryland

Testimony

W.S. Sellman, Office of the Assistant Secretary of Defense, Force Management and Personnel, Washington, DC

Walter Hathaway, Portland Public Schools, Portland, Oregon

Wayne J. Camara, American Psychological Association, Washington, DC

Jon Magoon, University of Delaware, Newark, Delaware

Mary E. Dilworth, American Association of Colleges for Teacher Education, Washington DC

William Gray, Lawyers Committee for Civil Rights under Law, Washington, DC (Speaking for FairTest)

Waveline T. Starnes, Office of Instruction and Program Development of Montgomery County Public Schools, Rockville, Maryland

Roland K. Yoshida, Queen's College, City University of New York, and

Anthony A. Cancelli, John Sowinski,
and Regis Bernhardt, Fordham
University, New York

***The Effects of Testing on American
Indians***

Co-sponsored by the Native
American Scholarship Fund, Inc.

Albuquerque, New Mexico

April, 21, 1989

Invited Speakers

Dean Chavers, President, Native
American Scholarship Fund, Inc.,
Albuquerque, New Mexico; and
Patricia Locke, Director, International
Native American Language Institute,
Mobridge, South Dakota: The Effects
of testing on American Indians

Panel Members

Rodney Brod, Professor of Sociology,
University of Montana, Billings,
Montana

Tony C.M. Lam, Department of
Educational Foundations, University
of New Mexico, Albuquerque, New
Mexico

Damian McShane, Associate
Professor, Utah State University,
Logan, Utah

Olivia Skenandore, Teacher, Laguna
Pueblo Elementary School, New
Mexico

Gender Bias in Standardized Testing

Co-Sponsored by the National
Women's Law Center

Washington DC

October 13, 1989

Invited Speakers

Esther E. Diamond, Educational and
Psychological Consultant, Evanston,
Illinois: Overview of the issues

Phyllis Rosser, The Equality in
Testing Project, Holmdel, New Jersey:
Gender differences in problem
solving

Sylvia T. Johnson, Department of
Psychoeducational Studies, School of
Education, Howard University,
Washington DC: Predicting the
performance of minority girls and
women from standardized tests.

Panel Members

Nancy Cole, Educational Testing
Service, Princeton, New Jersey

Carol Nagy Jacklin, Department of
Psychology and Program for the Study
of Women and Men in Society,
University of Southern California,
Los Angeles, California

Isabelle Katz Pinzler, Women's
Rights Project, American Civil
Liberties Union, New York

Cynthia Schuman, National Center
for Fair and Open Testing (FairTest)
Cambridge, Massachusetts

Linda Winfield, Center for Research
on Effective Schooling for
Disadvantaged Students, The Johns
Hopkins University, Baltimore,
Maryland

Leslie R. Wolfe, Center for Women's
Policy Studies, Washington DC

Appendix:

Invited Papers

THE FOLLOWING INVITED PAPERS have been published in *Test policy and the politics of opportunity allocation: The workplace and the law*, B.R. Gifford, Ed. (Boston: Kluwer Academic Publishers, 1989).

Adams, R. F. Economic models of

discrimination, testing, and public policy.

Butler, J.S. Test scores and evaluation: The military as data.

Chachkin, N.J. Testing in elementary and secondary schools: Can misuse be avoided?

Fisher, R.C. Los Angeles testing policies: Conference remarks.

Fremer, J.J. Testing companies, trends and policy issues: A current view from the testing industry.

Gelerter, R. The uniform guidelines and subjective selection criteria and procedures: Conference remarks.

Gifford, B.R. The allocation of opportunities and the politics of testing: A policy analytic perspective.

Levin, H.M. Ability testing for job selection: Are the economic claims justified?

Neiner, A.G., & Love, W.D. Examples of testing programs in the insurance industry and a discussion of employment testing policy issues.

Patterson, P. O. Employment testing and Title VII of the Civil Rights Act of 1964.

Rebell, M.A. Testing, public policy, and the courts.

Schwartz, D.J. Non-discriminatory use of personnel tests: Conference remarks.

Webber, C. The mandarin mentality: Civil service and university admission tests in Europe and Asia.

The following papers have been published in *Test policy and test performance: Education, language, and culture*, B.R. Gifford, Ed. (Boston: Kluwer Academic Publishers, 1989).

Baker, E. Mandated tests: Educational reform or quality indicator?

Haertel, E. Student achievement tests as tools of educational policy: Practices and consequences.

Haney, W. Making sense of school testing.

Hanford, G. Advice to the Commission: Conference remarks.

Jencks, C. If not tests, then what? Conference remarks.

Kochman, T. Black and white cultural styles in pluralistic perspective.

Madaus, G. The Irish study revisited.

Mehrens, W. A. Using test scores for decision making.

Moore, E.G.J. Ethnic group differences in the Armed Services Vocational Aptitude Battery (ASVAB) performance of American Youth: Implications for career prospects.

O'Connor, M.C. Aspects of differential performance by minorities on standardized tests: Linguistic and sociocultural factors.

Tsang, C. L. Informal assessment of Asian Americans: A cultural and linguistic mismatch?

Valdes, G. Testing bilingual proficiency for specialized occupations: Issues and implications.

The following papers will appear in a third volume in the NCTPP series:

Test policy in defense: Lessons from the military for education, training, and employment, B. R. Gifford and L.C. Wing, Eds. (Boston: Kluwer Academic Publishers, in press).

Eitelberg, M.J. Becoming brass: Issues in the testing, recruiting and selecting of military officers.

Sticht, T.G. Military testing & public policy: Selected studies of lower aptitude personnel.

Wise, L.L. The validity of test scores for selecting and classifying enlisted recruits.

The following book-length manuscript also will be made available in the Kluwer series:

Haney, W., Madaus, G., & Lyons, R. The fractured marketplace for standardized testing (Boston: Kluwer Academic Publishers, in press).

The following papers will appear in a forthcoming volume on cognitive approaches to assessment, to be edited by B. Gifford and M.C. O'Connor and published by Kluwer.

Brown, A. Campione, J., Webber, L., & McGilley, K. Interactive learning environments: A new look at assessment and instruction.

Gardner, H. Assessment in context: The alternative to standardized testing.

Resnick, L.B. & Resnick, D. Assessing the thinking curriculum: New tools for educational reform.

Sternberg, R.J. CAT: A program for complete abilities testing.

Remaining papers invited by the Commission will be published by Kluwer in forthcoming volumes or made available through the ERIC Clearinghouse on Urban Education (CUE), Teachers College, Columbia University, or published elsewhere.

Bishop, J. The economics of employment testing.

Butler, E.P. & Lawrence, B.N. Testing and assessment in publicly-supported job training for the disadvantaged.

Chavers, D. & Locke, P. The effects of testing on American Indians.

Clune, W. Courts as cautious watchdogs: Constitutional and policy issues standardized testing in education.

Cunningham, A. Eeny, meeny, miny, moe: Testing policy and practice in early childhood.

Duran, R. Testing of Hispanic students: Implications for secondary education.

Goldstein, H., & Wolf, A. Recent trend in assessment: England and Wales.

Haney, C. & Hurtado, A. Standardized error: Testing and employment discrimination.

James, L.R., Demareo, R.G., & Mulaik, S. A. Critique of validity generalization research.

Jensen, A. A Psychometric G and manifest achievement.

Jones, L.V. School achievement trends for black students.

- McDowell, D.S. & Dodge, G.E.
Employee selecting and Title VII
of the 1964 Civil Rights Act:
The legal debate surrounding
select criteria validation and
affirmative action.
- Mehrens, W.A. Issues in teacher
competency.
- Pennock-Roman, M. The status of
research on Scholastic Aptitude
Test (SAT) and Hispanic
students in postsecondary
education.
- Peters, C.W., Wixon, K.K. Valencia, S.,
& Pearson, P.D. Changing
statewide reading assessment:
A case study of Michigan and
Illinois.
- Reilly, R., & Warech, M.A. The
validity and fairness of
alternatives to cognitive tests.
- Rosser, P. Gender and testing.
- Smith, K.U. Human-factors analysis of
employment.
- Valdés, G., & Figueroa, R. The nature
of bilingualism and the nature
of testing: towards the
development of a coherent
research agenda.
- Wiesen, J.P., Abrams, N., & McAtee,
S.A., Employment testing: A
public sector viewpoint.
- Wilson, M. Moore, S. & Gumpel, T.
The assessment of validity for
selection rules: Evidence for
linearity assumption and
implications of its failure.