THRESHOLDS

IN EDUCATION

Vol. VI, No. 1 1980

Standardized Testing

In Memoriam: Oscar K. Buros
Trends for the 80's
Criticisms Discussed

Yes, I would like to have your quarterly magazine.

To: Editor, **Thresholds in Education**
P.O. Box 771
DeKalb, Illinois 60115

Please enter my subscription for:

☐ one year	$ 8.00	☐ payment enclosed
☐ two years	$15.00	☐ bill me
☐ three years	$21.00	

Ship to: _____
_____
_____

THRESHOLDS IN EDUCATION

Vol. VI No. 1
1980

Thresholds is entered as Third Class mail at the Post Office in DeKalb, Illinois under permit number 265 and under provisions of Part 144, Postal Manual Authorization for mailing as a non-profit organization was secured on February 18, 1975.

Subscription Information. Subscription rates are as follows: one year $8.00, two years $15.00, three years $21.00. For foreign subscriptions other than Canadian add $2.00 more per year. Send to: Editor, **Thresholds in Education,** P.O. Box 771, DeKalb, IL 60115.

Thresholds is published quarterly in February, May, August and November.

Why This Issue of Thresholds in Education Was Not Called "Standardized Testing: Critical Issues"

Your editors initially had a cover designed with "Critical Issues" prominently emblazoned at the bottom. Then is occurred to us that we had succumbed to the modern journalistic urge to shout. In an age of catastrophes and near catastrophes how can the claim be made that anything about standardized testing is truly a "critical issue?"

The issues involved in standardized testing more closely resemble toothaches than emergency appendectomies. The calamities of the moment will of necessity subside because of their urgency; the toothaches—but please permit me to abandon the analogy.

The selective perceptions of an informed citizenry still permit attention to fine-tuning of human experience. We regard it as remarkable that the survivors of inflation, unemployment, crime, TV, national humiliations, energy shortage, cancer, youth, burnout, international hooliganism, and American politics (to name a few disconcerting aspects of this society) can still engage in, or follow, the debate over the uses and misuses of standardized tests.

However, even with the demands upon our attention and energies from the outside, persons whose private and professional lives are influenced by tests are encouraged to continue to examine (critically?) the test instruments and their applications. Some students of testing have succeeded in narrowing the scope of their concern to certain populations and even to specific test items. Others have helped in creating new measuring devices or to perfecting older instruments. Still others have taken up causes or promoted interest in new or different types of devices or uses.

Notable among interest groups which have studied and/or criticized testing are the National Education Association and the American Psychological Association. The report from Ralph Nader's group warrants mention as well as the statements of the Educational Testing Service and Arthur R. Jensen's lengthy statement in his book **Bias in Mental Testing.** Again, we fear that we have strayed in venturing to name a few and neglecting the many voices for standardized testing. We trust that "letters to the editor" will help to give vent to the fervor of those whose ideas were not solicited in time to be included in this issue.

We know that pending legislation, usage, and court decisions may, at this moment, be shaping the course of standardized testing at various governmental levels. We also acknowledge acrimonious exchanges among groups with varied interests and concerns. Let us examine some of these issues and problems of standardized testing as they are currently found.

A last word. If there is a hero of standardized testing, it would have to be Oscar K. Buros. His work is his monument. As a sign of the respect which thousands of us feel, we asked Dr. David Tiedeman to write the lead article about Oscar K. Buros. We think Dave has succeeded in summarizing the greatness of the contributions of Buros and of the successful attempt of Mrs. Buros in finding a home for the Buros Institute of Mental Measurements at the University of Nebraska.

I wish I could say to him, "Oscar, I knew you only by your works, but I shall admire you forever."

Leonard L. Pourchot

ARE THE SOVIETS DEVELOPING PROHIBITED BIOLOGICAL WEAPONS?

Tax

red tape

hearing loss

politics

HOW TO STOP IN- FLATION

Oil prices.

the Soviet arsenal

AUTO INSURANCE

U.S. Oil Production

SAFETY STANDARDS

BUDGET

Arabs

Can Capitalism Survive?

Fiesta is assembled by Ford in Germany

Fear

Low productivity

Boycott

ENVIRONMENT

the recession began last year.

the twisted wreckage and the charred bodies of eight fallen American commandos . . .

"excessive" and "obscene" business profits.

Inflation Fight

Catastrophes

WHAT IS A "CRITICAL ISSUE"?

AVE BIG $$$ ON YOUR GROCERY BILLS

12½% FINANCING—29 YEARS—

"There is no man to follow Tito. He is unique."

FUSION ENERGY

the Hostage Crisis

WHAT IS GERM WARFARE?

DATSUN WE ARE DRIVEN

NISSAN

energy conservation.

THE ECONOMY

FOREIGN POLICY

The Economy: Sagging Fast

MAJOR RECESSION NOW FORECAST BY HILL BUDGET OFFICE

The Task of Dealing With the Russians

Cuba's Exodus

Less than 1 mg. tar, 0.1 mg. nicotine.

RETIREMENT

coal

Warning: The Surgeon General H- That Cigarette Smoking Is Dan-

Hostages

THE JOBLESS JOLT

Unemployment, the most visible mark of a recession, soared in April, putting new obstacles in the way of a balanced budget.

worldwide crisis

The Shaky Alliance

bureaucracy

CUTTY SARK

failure

The Meltdown

It has 17,000 employees and $14 billion to spend.

MISSILES

Death

despair

BUY ME!

It's time we all went on an energy diet.

bankruptcy

DEGENERATE

mistakes

Retirement Pitfalls

In Memoriam
OSCAR KRISEN BUROS
1906-1978

by David V. Tiedeman

I

Oscar Krisen Buros is dead. Long live The Mental Measurements Yearbook!

Logically enough, Oscar and Luella Buros issued the 1938 Mental Measurements Yearbook in 1938. But the 1938 Yearbook subsequently became known as the First Yearbook as the intrepid Buros scholars acknowledged in their 1949 publication of the Third Mental Measurements Yearbook both that the annual publication proved to be inadvisable and that sequential rather than yearly designation in title proved to be preferable. Mental Measurements Yearbooks subsequently settled into a 3-5 year cycle.

The Buros's published seven other Yearbooks in the forty years since publication of their First Yearbook. Their last Yearbook, the Eighth Mental Measurements Yearbook appeared in November, 1979.

For all practical purposes, the Eighth Mental Measurements Yearbook was fully edited when Oscar Buros died on March 26, 1978. Some publication tasks remained after Buros' death. Luella Buros and her able Yearbook staff then steadfastly put their shoulders to the remaining publication wheel and the Eighth Yearbook appeared about on schedule as a result. This is but another in a long list of debts our society owes the Buros's and their staffs for forty-five years of duty at King Oscar's Round Table for Examining Tests.

II

Now is a critical moment to reflect both upon what The Mental Measurements Yearbooks offer society and upon what the times and society will demand of the Yearbooks' new editor.

The introduction of the Eighth Mental Measurements Yearbook offers a good starting place for reflection upon the critical transition now underway in the Yearbooks' editorship. The Eighth Yearbook's Introduction lacks crusading Oscar Buros' typical punch because it is written posthumously. However, that Yearbook's introduction continues in Oscar Buros' inimitable style and grace because Luella Buros and her staff merely updated his Introduction to the Seventh Yearbook. By 1972, Buros had the essentials of the enterprise to which he and Luella Buros dedicated their careers firmly in mind. Buros made the Yearbooks part of an information system whose published elements (1) originate upon publication of a test anywhere in the English speaking world and (2) exist until the better of the published tests are (a) categorized, (b) indexed, (c) individually identified by their salient characteristics and associated with frankly critical balanced reviews and a bibliography of verified studies in which they have figured, and (d) run the course of their usefulness.

During the forty-five years in which the Yearbooks have been conceived and published, the Buros's developed what is unquestionably the most complete library of tests in the English speaking world. They also developed a bibliographic resource now numbering over 77,000 entries. In addition, the Eighth Yearbook lists 576 books on testing and assessment techniques as well. This list and those published in prior documents identify practically all the books on testing and assessment techniques written in English and published since 1933. Finally, the Buros's derived over the years a system of test information publication in which the MMY stands as core and is supported by two volumes of Tests in Print as well as by monographs of critical review in several specific subject areas. The Tests in Print series permits the Buros's to publish more exhaustive test bibliographies than is possible in what has now become the two volumes of the Eighth Yearbook. And the monographs permit publication of additional reviews of some tests and the inclusion of tests in a subject area which are excluded from publication in a Yearbook because of the intense competition for space which arises when a topic becomes a part of the larger corpus of tests published in the English speaking world.

No Mental Measurements Yearbook could have been published nor could have flourished without the cooperation of test publishers, professionals voluntarily reviewing tests critically, test users with need to know test qualities, and Oscar Buros leading the crusade for better tests and testing. Oscar Buros steadfastly believed that tests can be constructed, marketed, criticized, and used with integrity. Oscar and Luella Buros staked their careers and their livelihoods on those beliefs. Buros proudly notes in the 7th MMY that the Yearbook was published without any outside support whatsoever. Furthermore, the Buros's published eight of the ten major publications then in the MMY system by themselves. The 8th Yearbook continues in that fine entrepreneurial tradition. But continuation of this tradition presently needs reassessment as the editing of the Yearbooks is transferred from the Buros's to an as yet to be named editor leading the newly formed Buros Institute of Mental Measurements at the University of Nebraska. [See footnote, Ed.]

III

Beginning in 1976 Oscar and Luella Buros invited persons and/or institutions to take up their challenging crusade of keeping the house of testing in order while serving as King Critic in that house. As one who, in collaboration with several colleagues, seriously tried to take up this challenge, I can attest that the cost of publishing a Yearbook is beyond my poor means and beyond the means of a state university such as mine which is restricted by law to manage its cash flow within a twelve-month period as well. It appears that the University of Nebraska has now solved the cash flow problem which Northern Illinois University and I could not solve.

Long live The Mental Measurements Yearbook then? It depends, it seems, upon the appearance of subventions and ingenuity at this juncture in life. Needed

David V. Tiedeman is Coordinating Officer for Vocational, Technical and Career Education at Northern Illinois University.

subventions must be of two kinds, investment in data base conversion to electronic data processing, and capital to underwrite publication of new books should it appear that this old form of data base printout ought to be continued. Ingenuity will be needed in inventing the needed system, converting the present data file, and keeping a market together. Ingenuity will also be needed to meet several social challenges of testing as well. Hence, subventions and ingenuity must be generated with several policy matters squarely in mind, policy matters best brought to the forefronts of our minds at today's critical juncture in the MMY history.

Oscar Buros lived and died in greatness. During his life he scrupulously maintained his independence as the English speaking world's chief test critic by refraining from accepting money for advising any test publisher about construction of a test. I remember with a smile, Buros' story about his conscience struggle during 1977 when the American College Testing Program honored him by inviting him to lecture on a topic of his choice and later to sit with ACT professionals to reminisce and speculate about tests. Buros finally took the challenge and gave us his wide-ranging paper on status and prospects in testing as a last momento of his thought on test policy (Buros, 1979). But when Buros adopted the role of King Critic of tests in the English speaking world he adopted it to the hilt. Replacing Buros with a person of similar neutrality will be a feat of no small proportion. Quite a few experts in mental measurements market tests themselves. They are therefore ineligible for MMY editorship should Buros' standards prevail in picking his successor. I once passed Buros' neutrality tests but only after detailed quizzing about my loyalties to the College Entrance Examination Board following my three years' service in the chair of its Commission on Tests at the end of the 1960's. Buros really set a high principle and engaged in hard tests of it. I trust that the Buros Institute of Mental Measurements can continue to exist under this principle of action based on conscience independent of self interest in some specific tests.

Oscar Buros died in greatness as he lived in greatness. Buros left explicit instructions with Luella Buros that he was to be cremated immediately upon his death and that we friends should be immediately invited to visit 220 Montgomery Street, Highland Park, New Jersey to honor the Institute of Mental Measurements, not him. Had you visited his Institute, you would have been additionally impressed by the idealist Buros as craftsman. Besides devoting 45 years of his career to the examination of testing, Buros built the best test collection in the English speaking world, housed it in his home, and published its Yearbooks and supporting publications from the same home which served as the Gryphon Press in its two lower floors. The Mental Measurements Yearbook is Oscar and Luella Buros in the true sense of being. However, we cannot let The Mental Measurements Yearbook remain just Oscar and Luella Buros. We must now together turn it into an organization whose purposes will survive and better its old leadership.

Not all institutions which can publish the MMY are the right institutions to do so. In the Introduction of the 7th MMY, Buros noted:

> Test publishing is characterized by two extremes. On the one hand, a few publishers are in a dominating position; the top three percent of the publishers by number of tests listed in this yearbook account for 37.2 percent of the 1,157 tests. On the other hand, the number of small publishers is extemely large; 55 percent of the publishers are one-test publishers accounting for only 11.5 percent of the tests. (p. xxxvii)

This condition continues to prevail and figures in the transfer of the Buros Institute of Mental Measurements to the University of Nebraska. A university had the necessary neutrality. Fortunately, the University of Nebraska believes it can handle the MMY cash flow and new investment problems as well.

Unfortunately, the transfer of the Buros Institute of Mental Measurements to the University of Nebraska takes place far enough after Buros' death and publication of the 8th MMY to result in possible danger to the Institute's most precious possession, its test and review data bases. As indicated above, the Buros Institute of Mental Measurements presently owns the most extensive collection of tests, books on tests and assessment, and citations of test studies anywhere in the English speaking world. But a collection which lies fallow for even as few as six months begins to decline. When additions are not made regularly, the record of what is in a collection and what is not loses clarity and precision because no active mind is at work with its order. When persons don't use the collection from day to day for professional purposes, recovery of materials gets harder, categories harden, and the value of the collection dwindles because there is no living recovery system available for consultation on use. For these reasons, the decision to continue the tradition of the MMY had to be made in a short time. Actually, the decision was made within eighteen months of Buros' death and it remains to be seen if the value of its principal instruments, its data bases, has dwindled to the point where inordinate expenditures become necessary to restore the collection to its present condition of excellence and ready use. At the least, catch up will probably be the name of the BIMM data base game for quite a while at the University of Nebraska.

The existence of the IMM data base and its centrality to the MMY concept makes me feel that the IMM and its MMY are now appropriately situated, namely as part of a University. If so inclined, a university can maintain collections as persons responsible for the collections change positions, retire or die. Furthermore, the fact that the IMM data base now used in MMY production has not yet been in an organization such as a university where computer systems are available and where work can immediately start on making the IMM data base a part of existing electronic information exchange systems presents another challenge to BIMM management perspicacity at the University of Nebraska. The computerization of the BIMM data bases and their maintenance on computer discs and/or tapes consitutes another potential cost needing consideration in the transfer of the IMM and its resources from the Buros' estate to the University of Nebraska. Nebraska has stood forth for test examination in public interest which requires healthy subventions if the task is to be completed prior to publication of a Ninth Mental Measurements Yearbook in 1982, the year in which the MMY ought to be published. Will the 9th MMY be in its old tradition or move into the tradition of an electronic data system queried as needed on a pay as you go basis?

IV

The University of Nebraska which has risen to lead King Oscar's Round Table for Examining Tests not alone must reflect informed neutrality toward test publishing and act quickly, it must also wrestle with some new social issues and technical advances not central during King Oscar's reign.

Today, tests stand naked before many courts of public and legal opinion. In many regards, tests are being tried for things of which some tests and their authors and publishers are not guilty. On the other hand, tests are presently being tried for many things which persons who considered themselves unfairly treated by tests insist that tests and their users did. The temper of these times, therefore, demands continuation of Buros' crusade for better tests, for better information about tests, and for better use of tests. Test equity gives every indication of emerging as a criterion for test evaluation which must in the future be treated on a par with the concepts of reliability and validity by which we professionally measure tests so assiduously at present.

Test equity will also change the conditions for sale of test criticism in the future. With test users now enjoined by

court orders to demonstrate the validity of their tests in the specific situations in which they are used for selection among job applicants, there will both be more validity information constructed on tests than heretofore and more secrecy surrounding their use. Low predictive validity coefficients are becoming more and more detested by a better and better informed public. Institutions which use tests in entry selection and career promotion are going to be watched pretty carefully in the future.

How can an MMY still dedicated to critical test reviews ameliorate such attitudes so that the larger quantity of valuable test data will come to professional, if not public, light? But more interesting still, what role can an MMY play in critical test selection when personnel uses of tests have to be justified by company specific rather than general population data? Paradoxically, irrationally pursued demands for test equity and important, not just significant, validity may put an additional burden on the socially desirable MMY just when its market is more uncertain during potential editorial transition.

Oscar Buros knew that the truth dims under secrecy. He, therefore, did all he could both to honor the need for secure tests and to see that they were criticized in balanced ways just as every other open test was criticized. Our field needs crusading editors of The Mental Measurements Yearbooks to continue to attack the inadequacies of tests and testing and to lift veils of secrecy from secure tests without injury to their security. Buros felt that this issue deserved very careful consideration in future conduct of the MMY enterprise.

By similar token, technological developments in testing and test interpretation raise new issues in critical review of test practice. Several tests are now being interpreted individually by means of computer systems. The algorithms and data on which these computer systems operate must become just as much a matter for public scrutiny as are tests and manuals in print. However, means will have to be found to bring such computer algorithms and data into public scrutiny and judgment without endangering their monopolization just as Oscar Buros found means to gain criticism of secure tests without endangering their security. The new editor of the MMY will, therefore, have to be wide-ranging in relation to modern information exchange if this issue is to find definition and resolution. The same stipulation holds for dealing critically with the new generations of criterion-referenced and competency-defined tests standards now coming into popular use.

Computing systems which can approximate professional test interpretation in individual cases manifest the technological side of a potential shift in the scientific paradigm on which test construction and interpretation may be based in the future. Criterion-referenced and compentency-defined standards manifest the personal meaning goal arising as a part of that potential shift in scientific paradigm which is now on the horizon. The potential shift in the scientific paradigm itself—a shift from a position in which a subject never participates in determining the meaning of that on which he or she is measured to one in which the subject collaborates with the scientific investigator in determining the psychological meaning of concepts and action for the subject until such time as judgment must ultimately become unilateral for the judge—looms large on the horizon at the moment. I made the potential shift the basis of my suggestions to the CEEB Commission on Tests in the late 1960's, (Tiedeman, 1970), but Leona Tyler provides a much more complete account of this potential shift in her 1978 book, (Tyler, 1978). Should our scientific paradigm shift to a life process paradigm as it now seems to be doing, test construction and criticism will have to shift to a testing as process paradigm as well. Tests will then be more parts of structured interview situations (the inquiry processes) in which the tester (sometimes even a computer system) and the subject proceed along a line of inquiry designed to lend hierarchical restructuring to the subject's comprehension of an individually and societally interesting facet of concepts and action. And test criticism will then have to incorporate review of the structured interview as well as the test situations in which the system puts the subject for mutual decision concerning the state of the subject's concepts at particular moments. These possible conditions may well boggle the mind but they out to stand as challenge to the new editor of the MMY because developments like these are on the market and in the marketing wings of the near future. MMY of the future owes its society critical judgment on such emerging processes and products as well as critical judgment about test processes and products continuing in the molds of today.

V

Oscar Krisen Buros is dead. Long live The Mental Measurements Yearbook? It will, now that the University of Nebraska has assumed that responsibility. However, the Yearbooks will live only if subventions materialize to enable the Buros Institute of Mental Measurements entrusted with continuation of Buros' data bases and supporting reporting systems to adapt new modes of information sharing now coming into being. In addition, the Yearbooks' new editor faces the big challenge of sensing the signs and currents of the times as they bear on tests, testing, and the making of money without violating public trust in test action pursued with professional integrity. May such virtuosity materialize in short order because countdown time on the future of MMY is in process. After all, Oscar Krisen Buros died two years ago now. A new apparatus is presently in place at the University of Nebraska. But that apparatus needs to get going fast.

Luella Buros lost a dear husband and a cherished colleague on March 26, 1978. Oscar Buros' staff lost the leader whom they had invaluably served with deep dedication. I lost a cherished colleague—a colleague from whom I learned much about the real meaning of work and of how to die. We professionals lost King Oscar of Our Round Table for Examining Tests. We momentarily grieve with Oscar's wife, staff and friends over our loss of a great man. However, King Oscar would rather we celebrate and continue what he stood for: examined test practice. It's time we got on with that task. Let's do it! The Buros Institute of Mental Measurements at the University of Nebraska needs the help of all of us in the field in this, its Moment of Truth.

References

1. Buros, Oscar Krisen. The Eighth Mental Measurements Yearbook. Highland Park, NJ: Gryphon Press, 1979. pp. 1972-1983.
2. Tiedeman, David V. "Can a machine admit an applicant to continuing education?" In Commission on Tests. Report of the Commission on Tests: II Briefs, New York: College Entrance Examination Board, 1970. pp. 161-182.
3. Tyler, Leona E. Individuality. San Francisco: Jossey-Bass, 1978.

Dr. Cecil R. Reynolds, Acting Director of the Buros Institute of Mental Measurements, announced in April that Dr. James Mitchell has accepted the position of Director of BIMM and will be starting in August, 1980. Ed.

Standardized Testing — Trends of the '80's

by Robert M Resnick

A variety of factors in the educational environment suggest that content validity and instructional management will be the two most critical measurement issues of the next decade. The test publisher will be impacted by these issues in terms of the important changes they may cause in the ways that educational tests are developed, configured, administered, and used.

Content Validity

CTB/McGraw-Hill has found a steadily increasing number of schools, districts, and states expressing a need for content-customized measures of student achievement. These educational agencies have determined, explicitly or more often implicitly, that they no longer need or want standardized tests to play an active role in cross-fertilizing school curricula. The intention of such customized tests is to measure attainment of a unique set of instructional objectives considered most appropriate by a particular school or district. In short, an instrument is required to test precisely what is taught.

While many districts have expressed this need to develop unique testing instruments, few have been able to afford the cost of hiring a test publisher and most do not have personnel who are proficient in the technical skills required to develop good tests. Even for those with the necessary funds, a further constraint has been that customizing a test has always precluded the possibility of providing norm-referenced scores.

This situation is likely to change dramatically in the '80s. Advances in Item Response or Latent Trait Theory and computerized typesetting and photocomposition should enable test publishers to provide affordable content-customized measures for which norm-referenced scores are available. New item calibration techniques, including the Rasch or One-parameter model and the more sophisticated Three-parameter Latent Trait Model, will make it feasible to create true item banks. These banks will consist of calibrated items together with the statistical and content information necessary for the construction of tests according to statistical and content specifications. Computerized assembly of test booklets will significantly improve the flexibility and cost effectiveness of test production procedures.

Item bank products will be special purpose tests built to match customers' curricular and measurement needs. By definition these will be content valid measures. Equated alternate forms will be readily available to those concerned about test security. Testing time can be minimized with computerized adaptive or tailored versions of the same basic special purpose test. The important fringe benefit derived from the use of alternate forms and adaptive tests is that they will encourage teaching to the objectives of the test since the specific items will change from one administration to the next in the case of alternate forms, and from one student to the next in the case of adaptive tests.

Instructional Management

Once it is resolved that a test is content valid and can be used as an instructional tool, the question becomes "how can testing be integrated with instruction?"

A distinction needs to be drawn between actual integration of tests with instruction, and linkage, where tests are used to route students into remedial or supplemental learning activities. Integration is a break through that has yet to be realized while linkage has been accomplished by a number of existing instructional management systems.

The problem with simple linkage is inefficiency. That is, the learner is required to continually move between disparate activities called "testing" and "instruction". With the advent of a technology for acquiring and analyzing student reponses and presenting appropriate instruction, it may not be too much longer before educational research provides the missing key to enable educators to use student responses to guide the instructional process signaling the true integration of testing and instruction.

Aside from the technological considerations already cited, there are instructional, economic, and other influences leading to the conclusion that instructional management will become increasingly important in the next decade. These influences include the following:

1. Instructional — Even the most vocal critics of the testing industry will usually agree that education needs some kind of testing to provide teachers with immediate helpful information about students. That is, to help teachers most effectively, tests should be diagnostic and should pinpoint children's problems closely enough to help guide further instruction. Implicit in the phrase "guide to further instruction" is a critical need for transportation from a diagnosis to some set of appropriate instructional activities.
2. Economic — Any complicated system must monitor the results of its services to find out how effective it is. Educational systems faced with negative demographics and uncertain funding are under steadily increasing pressure to control costs and to demonstrate improved service to the school population. An integrated diagnostic/prescriptive system holds the promise of increasing the productivity of each classroom by directing the instructional focus to the areas of greatest need. The expense of instructional materials can also be reduced by selectively purchasing materials based on a review of the needs of each classroom. Improved educational service is also likely to occur due to the emphasis the schools can place on direct instruction intended to remediate areas of need.
3. Theoretical — As Bllom and others have pointed out, the development of feedback and corrective procedures at various stages of the learning process is central to "mastery learning". It is more than likely that "time" will continue to

Robert M. Resnick is publisher, California Test Bureau/McGraw-Hill.

be regarded as a critical variable in school learning in the next decade and beyond, and it will be considered essential that an instructional management system provide quick transportation from diagnosis to "time on task".

4. Technological — The concept of supplementing a teacher's work with computing power has been difficult to implement using classroom terminals and main-frame computers. However, the recent development of relatively inexpensive microcomputers has led to an explosion of interest in hardware/software systems at the school level. The imminent availability of this electronic technology will make it feasible, from a time and an economic standpoint, to manage individualized instruction without burdening the teacher with excessive paperwork and an information overload.
5. Social — Continued emphasis on fair treatment of individuals regardless of race, sex, language, and/or exceptionality will reinforce the philosophy that most students, given favorable learning conditions, can become very similar with regard to learning ability. Paradoxically, one of the necessary conditions for encouraging equal learning is likely to be an allowance for individuality in thinking and learning styles. The obvious way to accomplish this objective is to always key and categorize a wide variety of instructional materials and activities to each unit of diagnosis and to have them readily available to each learner.
6. Legal — The mandates of equal protection and due process will require that any form of competency testing provide for immediate feedback and timely remediation in those cases where it is necessary. Teachers will need more instructional support than is currently provided in the majority of states as the notion of "instructional validity" is tested in the competency litigation which is anticipated in the next decade.

Although major achievement tests have always intended to reflect, not determine, curriculum in the schools, it is evident that this will not be sufficient in the future. As schools take a more active role in determining the framework for curriculum and instruction, test publishers will be expected to provide the broad range of services required by a customizing environment. Should both the schools and the publishers succeed, it will be the nation's students, teachers, and parents who are the beneficiaries. The convergence of these goals should put more relevant and more usable information in the hands of teachers and administrators, enabling them to better manage and improve instruction in our schools.

"Although major achievement tests have always intended to reflect, not determine, curriculum in the schools, it is evident that this will not be sufficient in the future."

Standardized Testing in the '80's

by Jerome E. Doppelt

There is a relatively new term in the language of education: "standardized testing." In an earlier incarnation "testing" alone seemed sufficient to identify a process of examination of people for a variety of purposes. But as we have grown and matured in the ways of measurement, the term "testing" seems to have acquired a qualifying adjective. "Standardized testing" now refers to the abundant number of professionally constructed measurement instruments which are carefully designed to meet stated purposes, which are to be administered and scored in prescribed ways, and for which comparative data such as norms have been collected.

One can, and many do, see this term as the starting point of a battle royal about the dangers and the values of attempts to measure the population's attributes. Some also see the process of standardized testing, as distinguished from teacher-made testing, as a means of making insidious comparisons among those who teach rather than among those who are taught. This involves a specter called "accountability" which, to many, is more horrendous than "standardized testing." For the present discussion, however, there is no need to charge the term "standardized testing" with emotion. It is used to describe a type of offering—professionally constructed measurement instruments intended to meet needs that are educationally meaningful. Whether standardized testing will grow or shrivel in the'80's will depend on how well the needs of education are met. We believe that testing will grow because there are indications of more effective meeting of needs than is presently the case.

There are obviously limitations to any form of measurement but that does not rule out the useful applications of what can be measured. Achievement test scores can be used to compare classes or school systems; this is a particular type of use and if there is reason to obtain such information, the standardized test will provide it. Survey achievement tests and the various types of diagnostic tests are also useful for instruction. For the area of instruction particularly, two developments of recent times will have great impact on the future. One of these is the computer; the other is measurement technology.

Even in its present state of the art, the computer can convey a vast amount of information concerning an individual's performance. Not only total and sub-total scores can be obtained and compared with relevant normative data, but information down to the item level can be provided when appropriate. It is reasonable to expect that during the next decade the communication of the results of standardized testing will be more extensive and less expensive. The tendency, already evident, to include narrative along with numerical scores will grow and will be broadened to include a variety of audiences: teachers, administrators, students, parents and such groups as government agencies and the press.

Regardless of its increasing sophistication, the computer cannot produce information that is fundamentally of higher quality than the input it receives. It would be a shame to devote high-powered computer facilities to the development of data from tests of low quality. Professional test making and the application of modern measurement theory must go along with the power of the computer to turn out huge quantitites of information. We are fortunate in seeing now increasing interest and activity in the development of more effective means of measurement: latent trait theory and its uses in developing and equating tests; studies of item characteristics and how they affect the meaning of scores; analysis of test content from the viewpoint of ethnic subgroups; the increasing realization of the complementary uses of objective-referenced and norm-referenced measures.

In effect, what we will see happening in the next decade is growth in the volume and type of information that will be provided from tests that are more finely-honed technical devices. Narratives will probably include not only descriptions of what was found out about the student or group but prescriptive measures that might be considered in order to achieve certain results. Some of these things are being done now but more sophisticated offerings may be expected in the future. Although the precision of measurement may be increased by the application of complex theory, it should be possible to keep the language of the reports simple and understandable.

The anticipated events of the '80's may be categorized under the heading of "increased communication." The storms about the use of tests in schools, for admission to institutions, and for selection for jobs are often generated by the feelings of some people that others will use the tests to the detriment of individuals or groups. Increased communication about test results and the possible value of such results for teaching, placement, selection and guidance should help reduce the emotional aura currently surrounding standardized testing. The finer tuning of the suggestions that may result from more technologically advanced testing procedures should help convey a greater feeling of value to the information provided.

It is, of course, obvious that information per se is not harmful. It is clearly the concern of many that the misuse of information could be harmful. Increased understanding of information will reduce the danger of abuse. Effective communication with all who have a need to know will also make clear the usefulness of information which is not necessarily exhaustive. The fact that all aspects of behavior are not accounted for does not vitiate the value of information that relates to some aspects.

Jerome E. Doppelt is with the Pyschological Corporation.

Critical Issues in Standardized Testing

by Robert L. Ebel

Standardized testing is currently under attack. Articles critical of testing appear frequently in the popular press. A number of books dedicated to stating the case against standardized tests have been published. Legislatures in some states have enacted laws prohibiting the use of standardized tests of intelligence. Several professional associations of educators have urged a moratorium on standardized testing until better tests are available, and teachers better trained to use them wisely.

One whose opinions are shaped by these criticisms might conclude that standardized tests have lost favor and are headed for extinction. But the use of standardized tests of educational achievement in the schools of the nation has not declined appreciably. Legislatures in a number of states, including some that have banned intelligence tests, have enacted requirements for additional testing in statewide programs of educational assessment. Tests provide the principal source of data for the national assessment of educational progress, sponsored by the Education Commission of the States. Thus, despite the attacks, standardized testing continues to flourish.

Nevertheless, the attacks are real and substantial. They ought not to be ignored, or dismissed casually, by specialists in educational measurement. The remainder of this paper will be devoted to a discussion of several issues that have developed out of interest in and concerns about standardized testing.

1. Should standardized tests be used to measure student achievement?

Indeed they should, for when so used, they help schools greatly to do what they were created to do and are expected to do, which is to facilitate pupil learning of useful knowledge. They do this in many ways:

1. By keeping the eyes of all concerned with the learning process fixed on the main target.
2. By revealing differences in the effectiveness of different institutional procedures, or different curricular arrangements.
3. By motivating teacher's efforts to teach well and pupil's efforts to learn well.
4. By recognizing and rewarding success in learning.
5. By making comparisons possible between pupils, between teachers, between schools, of the outcome of their efforts to learn.
6. By causing schooling to become a purposeful educational enterprise whose results can be assessed systematically.

No school can do a good job and show that it is doing a good job without systematically auditing the results it is getting. Standardized tests of educational achievement provide one good means for making such audits.

2. Would a moratorium on standardized testing serve a useful purpose?

It would not. On the contrary, it would limit further our already limited efforts to assess the effectiveness of our educational efforts. It would remove the pressure that usage exerts on test makers to make a better product that will be more widely used. There is no reason to believe that a wholesale prohibition of continued use of the best tests we now know how to make would lead quickly to the production of better tests.

A moratorium is a legally authorized delay in the performance of an obligation. The obligation in this case is assessment of pupil achievement in learning, as accurately and meaningfully as possible. The delay proposed is in the use of carefully constructed and calibrated external tests of achievement. The purpose of the delay, presumably, is to enable the test makers to come up with better tests.

Those who propose a moratorium on standardized testing have offered no suggestions about the nature of the "better" standardized tests that they hope will replace the current "defective" tests. They have suggested no program for developing better tests. There is reason to suspect that they want not just a moratorium, but a permanent cessation in the use of standardized tests. They may believe that since the public might accept a moratorium, but not complete abolition, it is strategically wise to ask for half a loaf now, in the hope of getting the other half later.

I believe the advocates of a moratorium are wrong. To interrupt the use of standardized tests would limit further our already too limited efforts to assess the effectiveness of our educational efforts. It would remove the incentive for test makers to make better tests that will be more widely used. It would protect and preserve mediocrity in teaching and learning.

3. Are standardized tests inherently "racist" in design?

A racist is one who believes that the primary determiner of human traits and capacities is race, and that racial differences produce an inherent superiority of a particular race. A race, in turn is a division of humankind possessing traits that are transmissable by descent, and sufficient to characterize it as a distinct human type. While a test is itself incapable of having a belief, it might, if constructed by a racist, reflect such beliefs and thus tend to promote racism.

There are three reasons why it is unlikely that any standardized test is racist in design. First, there are by no means as many racists in our society as the frequent use of the term might suggest. Often the term is used not to describe but to condemn. It is used as an epithet to express resentment and hostility. Those who choose their words carefully seldom refer to another as "racist," because racist beliefs have been pretty thoroughly discredited. Thus, there are not many real racists that might be available to design tests.

Second, the roles of racist and test designer are somewhat incompatible. Racism is most common among the poorly educated. Test designers come mainly from the well educated. Third, standardized tests are designed to sell, and an obviously racist test is unlikely to sell very well in today's cultural climate. No sensible test designer would set out deliberately to reveal racist biases in a test.

The charges of racial bias in standardized tests have their origin in evidence that on the average members of one race make higher scores on the test than members of another. But the most reasonable explanation for this is that members of the one race have, on the average, developed more of the ability required by the test than have members of the other race. To say that these tests reflect middle class values, which is true in so far as they reflect any values, is misleading. For the tests are not measures of a person's values. They are tests of universally valid and generally useful knowledge, knowledge that is just as true for one race as for another, and just as useful.

Children may be born with equal rights, but they are not born with equal physical endowments or equal temperaments. Neither are they born into equally favorable environments for learning. Thus they quickly become unequal in abilities. The differences standardized tests report are very largely real differences in the developed abilities measured by the tests. To blame these differences on test bias is to blame the messenger for the bad news he sometimes

Robert L. Ebel is a noted author and authority in the field of Standardized Testing. Michigan State University.

brings.

4. How can tests be used to promote learning?

There is general agreement that the ultimate justification for the use of tests in schools is to facilitate learning. But there is a difference of opinion on what kind of use is most likely to enhance the learning desired. Some say that tests ought to be used more to help teachers to teach better, and less to judge them. They emphasize the use of tests for diagnosis and remediation. They question the use of tests to assess achievement in learning.

Others, and I am among them, believe that to neglect summative evaluation, the measurement of what has been learned is to give up the strongest influence for good that testing can have on study and teaching. That we have been guilty of that neglect is no small part of the reason for declines in school achievement in recent decades. Testing to help pupils to learn is not an alternative to testing for the purpose of determining how much they have learned. The two purposes are strongly supplementary, if not identical, in many school situations. If tests are used to measure what pupils have learned it will help them to learn by motivating and rewarding their efforts to learn. It will also motivate and reward the efforts of their teachers to teach.

Accordingly, I invite your attention to the following radical proposition.

Almost all teachers should be held responsible for providing evidence on how much their pupils have learned of what was set out to be learned.

It is radical in two senses. It is "marked by a considerable departure from the usual or traditional." It also goes to the "root" of the problem of excellence in education. Radical though the proposition may seem to be, it is nevertheless wholly reasonable. If we seek excellence in learning there is no better way to use tests to achieve that excellence.

5. What kind of tests should be used to guide college admissions?

Many people regard tests of innate capacity of learning as nearly ideal for college admission and scholarship programs. The term innate mental capacity suggests something which is fundamental and permanent, and hence well worth measuring, a divinity that shapes our ends, no matter how poverty, inadequate schools, or youthful follies may have rough-hewed them. A test of innate capacity presumably will not handicap the bright youth who grew up on the wrong side of the tracks. In theory, scores on it should not be affected appreciably by coaching, or indeed by instruction of any sort. Such a test would place no restrictions upon the secondary school curriculum. Local schools and individual teachers could presumably retain their freedom to teach what they choose, and it would make no difference if they taught it well or badly.

Even if tests to innate capacity for learning were available, they probably would not be desirable for college admission and scholarship programs. Innate capacity is not directly relevant to ability to profit from college instruction. Unless this innate capacity has been developed--unless the student has acquired considerable knowledge and numerous abilities—he is unprepared for college no matter how large his empty innate capacity may have been.

Thus the ideal test for use in guidance toward admission to college is a test of the pupil's achievement in having developed a broad and firm cognitive foundation for future learning. Such tests are likely to select the students best able to benefit from a college education. They are also likely to have stimulating, constructive effects on the programs of instruction in secondary schools.

6. Do criticisms of mental tests reflect flaws or abuses?

No doubt there are flaws in mental tests. No doubt they are sometimes abused. But the alleged flaws and abuses are often only pretexts. They are cited more often to condemn the tests than to show how they or their uses might be improved. In many cases the criticisms may be motivated by fear of what the tests could reveal, by fear of public reaction to evidences of unsatisfactory school achievement. Such criticisms do not deserve to be taken seriously.

To be destructively critical, to call for abandonment of an instrument or a practice, one needs little special competence. Critics who fire broadsides at tests and testing need very little knowledge of the principles and practices, the problems and their partial solutions. Constructive criticism is another matter. It calls for broad experience and thorough understanding. That is the kind of criticism tests have received for years in the Mental Measurements Yearbooks. That is the kind of criticism that deserves to be, and usually is, taken seriously.

7. Can standardized test scores contribute to the individualization of instruction?

Standardized tests reveal differences in ability. Individualized instruction ought to take account of, and be adapted to these differences. If it does, the tests can contribute to individualization. But there are two reasons why the educational value of this use of standardized test scores is not likely to be large. One is that standardized tests provide measures of general ability, whereas what is needed to individualize instruction are indications of the presence or absence of very particular abilities.

A second is that individualized instruction tends to be much less efficient, and thus much less effective overall than group instruction. Surely it is true that all learning is individual learning, but it does not follow from this that the instruction must be all, or even mainly, individual instruction. Teaching requires presentation, which can be done very well to groups of pupils. Learning requires assimilation by the pupil of what has been presented. Assimilation is an active process in which the pupil is the principal actor. She and she alone can adapt what has been presented to her own structure of knowledge. Of course, a teacher can help by answering individual questions or giving individual suggestions. But it would be a grave mistake to assume that the success of teaching depends mainly on how well the teacher adapts her presentations to the individual congnitive structures and learning styles of each different pupil. There is no credible evidence that a group of pupils learns more when instructed individually than when instructed as a group.

8. To what extent are tests themselves responsible for the lower scores that minorities tend to make on them?

A person's score on a mental test is the result of an interaction between the demands of the tasks that make up the test and the abilities of the person being tested. Thus the test itself is partly responsible for the score anyone makes on it, whether that person is a member of a minority group or not, whether the score is high or low. But should the test be blamed for the low score anyone makes on it?

Standardized tests are sometimes alleged to be biased against minorities. to the extent that they reflect elements of common culture they are biased in favor of that common culture and against different sub-cultures. There are indeed many sub-cultures imbedded in our common culture. But there are very few individual members of a sub-cultural whose lives are not guided, protected and rewarded almost entirely by elements of our common culture; our common schools, our common laws and government services, our common means of transportation and communication, our common sources of goods and services and jobs. If standardized tests reflect the common culture, nothing more should be asked of them. They can not, and indeed ought not to be asked to, reflect the distinctive features of a variety of sub-cultures. The best hope for those who receive lower test scores is not to change the tests. It is to acquire more of the abilities the tests require.

Educational Testing Service Responds to the Nader/Nairn Report

by Clyde Leib

A recent report on Educational Testing Service, written by Allan Nairn and sponsored by Ralph Nader, has charged that the major admissions tests administered by ETS on behalf of associations of colleges and graduate and professional schools constitute a "respectable fraud." The report asserts that exaggerated claims have been made for the tests, that as a result the tests have undue influence on admissions to higher education, and that in fact the tests have little value in predicting students' future academic performance.

The following is an authorized, condensed version of Test Use and Validity and Tests Scores and Family Income, written by Rex Jackson and published by ETS, which contain detailed responses to charges put forth in the Nader/Nairn report. Readers are encouraged to request free copies of these publications from: Information Division, Educational Testing Service, Princeton, NJ 08541.

* * * * * * * * * * * * * * * * * * *

The assertion made in **The Reign of ETS** that admissions tests have little predictive value will come as a surprise to the colleges and universities that have used scores on these tests as one element in admissions for decades. The faculty and admissions officers at these institutions know what the tests are designed to do and how they are used. Moreover, they have vast experience in selecting students for admission, in teaching these students, and in observing their academic performance. Apart from this experience, colleges and universities have conducted well over one thousand formal research studies of the relationship of test scores to the grades later obtained by enrolled students. Many of these studies have been conducted in collaboration with ETS, others by the institutions independently. **These studies show clearly that scores on the tests are useful in predicting the grades that students will obtain.**

The Nairn report arrives at its remarkable conclusions by misrepresenting the purposes and uses of the tests and by distorting or ignoring the results of research. Despite the fact that accurate information on these matters has been widely published and available to test users and to the public for a long time, the attention that the Nairn report will attract through its exaggerated and sensational charges requires that the record be set straight.

What the Tests Are Designed to Do

The major admissions tests administered by ETS are sponsored and controlled by associations of colleges and graduate and professional schools that use scores on the tests as one factor in admissions decision-making. Because grades and other information used in admissions (such as recommendations) are not directly comparable from student to student these institutions ask that applicants also submit test scores, which supplement the other information and provide a more uniform basis for evaluation of students' academic abilities and achievement. The institutions have long recognized the advantages to themselves in agreeing to accept results from uniform systems of professionally developed examinations and the advantages to their applicants in having results from a single examination, administered nationwide, acceptable for admissions purposes at multiple institutions. In fact, recognition of the benefits to applicants came first; the College Board was founded in 1900 by a group of colleges in response to the problems of secondary schools and their students in meeting different and uncoordinated college admission testing requirements.

The tests themselves vary in their characteristics. The major testing program at the college level, the College Board Admissions Testing Program, offers a test of scholastic aptitude (the SAT), a test of English usage designed primarily for decision-making about course placement of entering students, and a series of subject matter Achievement Tests. The Graduate Record Examinations (GRE) program similarly offers an aptitude test and achievement tests in areas of graduate study. Law schools and graduate management schools generally use tests that may be broadly characterized as aptitude tests—the Law School Admission Test (LSAT) and the Graduate Management Admission Test (GMAT).

Since the tests that are described as aptitude tests are most widely used, most attention focuses on them. A common misconception is that these tests somehow measure innate, unchanging abilities. In fact, they measure learned skills. They are described as aptitude tests because they are not tied to a particular course of study, curriculum or program, and because they are typically used to assess student's relative abilities to perform well in future academic work. Scholastic aptitude tests are most often made up of problems that test reading comprehension, verbal reasoning and vocabulary, mathematical reasoning, and data interpretation. They measure intellectual skills that students are expected to have developed through both school and non-school experiences, apart from the particular courses of study they may have pursued. These skills are also rightfully regarded as broadly applicable to success in a variety of future courses of study, since higher education at all levels requires ability to read with understanding, to reason with words and numbers, and to calculate and understand quantitative relations.

The philosophy underlying much of

education in the United States today places great stress on providing students with a broad base of general education and on delay of academic or professional specialization until a relatively late age. Students throughout high school and college are encouraged to acquire a broad acquaintance with the liberal arts and science. Considerable emphasis is placed on exploration, flexibility, and choice. There is no national syllabus. In this context, using tests of developed abilities, rather than relying exclusively on subject matter achievement tests at points of transition from school to college or from college to graduate school, is a consistent, appropriate, and sound policy.

Each of the major admissions tests is made up of multiple-choice questions. There are principally three reasons for this. **First,** use of these questions permits a broad sampling of problems of different kinds in limited testing time. Other forms of assessment cannot cover as much material in any given time, greatly increasing the risk that students will be evaluated in terms of only a few topics that do not fairly represent what they know. **Second,** many studies have shown that evaluation of essays or other free response exercises is fraught with unreliability. In practice, each paper can be read by only a few individuals, and the grades assigned by different readers to any single paper tend to vary a great deal. This also poses risks to students, since their scores will be affected by the standards and points of view of the particular readers to whom their papers happen to be assigned. In contrast, the questions on well-developed, multiple choice tests are reviewed by many individuals, and there is a consensus on the correct answers; scores on tests consisting of a number of such questions tend to be much more reliable than scores on essay tests. **Third,** multiple-choice testing and the automated scoring it permits have a number of practical advantages. The low costs of the major testing programs to students and timely reporting of information are, in large part, a result of this form of testing.

In short, the major admissions tests are designed to provide a common basis for evaluation, supplementary to students' academic records which vary in meaning from school to school; a reliable and fair means of assessing relative student abilities to perform well in future academic work, one which is not heavily dependent on the specific nature of students' previous academic preparation; and a means for students to satisfy admissions testing requirements of institutions across the country through taking a single examination at nearby testing centers and at relatively low cost.

Use of Tests in Admissions

Considerable space is given in the Nairn report to allegations that exaggerated claims are made for ETS tests—allegations that are supported through a few selected misquotations and quotations out-of-context from materials published about the tests. The result is an extraordinary misrepresentation of the facts. Booklets for students and users of scores on the admissions tests emphasize that the tests measure particular academic skills that scores are not infallible guides, and that the tests are designed to provide information to supplement students' prior school records and other evidence of academic competence.

Consistent with its misrepresentation of claims made for the tests, the Nairn report implies that scores on admissions tests are given undue weight, and that ETS is somehow an arbiter of admissions or "gatekeeper" to higher education. In fact, colleges and graduate and professional schools make their own admissions decisions, each institution using its own criteria. The criteria are decided upon by faculty and administrators at these institutions and in the case of many public institutions by state boards of education as well. The notion that ETS decides admissions is clearly a fallacy.

ETS and the sponsors of the admissions tests encourage institutions to weigh various kinds of information in admissions decision-making. But what are the facts about test use? Despite the advice institutions receive, do they generally place an undue weight on tests scores?

Most colleges are not selective and admit a large proportion of their applicants. A very large majority of students intending to go to college do so. A 1972 study conducted for the federal government found that of high school graduates who applied to college, 87.5% had been admitted to at least one institution by the end of their senior year (Hilton and Rhett, 1973). A 1978 survey conducted for the American Council on Education indicated that 75% of freshmen were attending their first choice college and nearly 95% were attending their first or second choice college (Astin, King, and Richardson, 1978, p. 18).

Even among selective colleges, test scores are seldom the most important factor in admissions, let alone the sole factor. A 1979 survey conducted jointly by the College Board and the American Association of Collegiate Registrars and Admissions Officers found that **under 2% pf the selective colleges responding to the survey indicated that test scores were "the most important factor" in admissions decision-making.** Ninety percent described test scores as a "very important factor" or as "one of several factors" (Van Dusen, Nelson, Jacobsen, and Ivens, 1979, p. 26). These general survey results are no cause for complacency about the issue of proper test use. But, they do suggest that test scores are much less cruicial in admissions decisions than Nairn supposes.

Evidence Supporting Test Use

Several kinds of evidence supporting use of the tests exist. At the foundation is an informed judgment by the designers and developers of the tests that the tasks sampled by the tests require skills that are important to competent academic performance—that the tests are in essence academic work samples. This judgment is guided by experience with particular types of questions or problems in other settings and by research showing a relationship between performance on these tasks and academic success. In addition, this judgment is subject to review of independent educators, particularly those in the institutions that use the tests. The practice of publishing sample test forms with answer keys for each of the major admissions testing programs further permits the widest possible scrutiny of test content. There is considerable agreement among those using the tests that the abilities assessed by the reading, verbal reasoning, and mathematical problems contained in the tests are relevant to successful academic work in their institutions. The prevailing view of the SAT is the one expressed by William Ambler, dean of admissions at Haverford College (as quoted in *Newsweek,* February 18, 1980): "The test reflects the words and symbols that students must deal with in courses every day."

Many technical studies of the tests are performed, including studies of test difficulty, the consistency of meaning of scores from one test form to another, the reliability (or stability and precision) of scores, and the relationships of scores on the tests to other variables. Studies of this last type are often concerned with the "validity" of a test—the extent to which scores relate to other measures of educational preparation, development, or achievement. Important among them are studies of predictive validity, or the extent to which test scores are related to future academic performance (usually measured by grades).

Typical results of prediction studies based on the test scores and grades of enrolled students are shown in Table 1.

Table 1

Characteristic Validity Co-efficients of Admissions Test Scores and Previous Grade Record (GPA) for Predicting Subsequent Grades

Admission Test	Type of School	No. of Studies	Median Validity Coefficients: Test Scores	Previous GPA	Both Predictors Combined
SAT	Undergraduate	827	.41	.52	.58
GRE	Graduate Arts & Sciences	24-30	.33	.31	.45
LSAT	Law	116	.36	.25	.45
GMAT	Graduate Management	67	.29	.21	.38

The numbers in the table are "correlation coefficients." A correlation coefficient is an index of relationship generally symbolized by the letter "r." An r of .00 indicates no relationship, (or more precisely no "linear" relationship). An r of 1.00 indicates a perfect relationship.

The numbers shown in the table are median values; actual values vary from institution. For a variety of reasons discussed in the full-length version of this paper, the numbers quoted in Table 1 are conservative estimates.

What are reasonable standards for test validities and to what extent do the major admissions tests meet these standards? Because ideal validity studies cannot be performed, the validities are not precisely known. Even without taking into account the various factors that attenuate validity coefficients, however, the obtained results indicate that the tests are useful predictors. A test with a validity of .40 is quite useful in identifying students who are likely to do well in academic course work at the college level.

The predictive validities of tests can also be judged in relation to those of the usual alternative predictors. At the college level, the characteristic validity of the SAT (.41) is not a great deal lower than that of high school grades (.52). In fact, these values vary from college to college, and in about 25% of the colleges the SAT is a better predictor than high school grades. At the graduate and professional levels, the tests most often have somewhat higher validities than does undergraduate GPA. **On the whole, the tests are about as useful in predicting future academic performance as previous academic grades, the predictor that is most widely used and accepted.**

Beyond Academic Selection

Nairn gives much attention to the issue of test validity, but he seems to recognize that his demonstration is unconvincing, for he also criticizes college grades, the criterion predicted by the SAT. Here again there seems to be a difference between Nairn's values and those of the colleges and universities, a difference far more fundamental than any dispute over predictive validity. Colleges clearly value academic talent and excellence as qualities to be sought and nurtured. Nairn's values are not precisely clear, but his report gives little attention to the importance of learning and academic achievement. This may help to explain why Nairn's views on how to go about admissions are so divergent from those of college and university faculties.

These comments are perhaps somewhat unfair to Nairn for probably he does share many of the values that his mentor, Ralph Nader, expresses in a preface to the report. There Nader indicates that he would prefer to give more attention to characteristics not measured by multiple-choice tests, such as "judgment, wisdom, experience, creativity, idealism, determination, or stamina." Elsewhere in the report, Nairn suggests that more weight should be given to non-academic activities and accomplishments.

Though few in the colleges would agree with abandonment of current academic measures, there is unquestionably much sentiment that other personal qualities like those cited by Nader should be given more systematic attention in admissions. Though these qualities are highly valued and sought, they are difficult to assess in any but the most subjective terms.

ETS and the client groups it serves have for many years funded, conducted, and published research designed to help colleges make better use of information on personal qualities in admissions. (See, for example, Anastasi, Meade, and Schneiders, 1960; Davis, 1964; and Baird, 1979; several of which are cited in the Nairn report). Currently the College Board and ETS are jointly engaged in a major research project in collaboration with nine colleges to develop improved assessments of personal qualities, special talents, skills, and significant accomplishments. As part of this project, the relationships of these measures to student development over the four college years—development broadly conceived in both academic and non-academic terms—will be studied.

ETS has no quarrel with Nader and Nairn's advancing their views concerning appropriate criteria for use in selective admissions. But, it cannot be agreed that Nader and Nairn's values are the only values, or that these values should be substituted for the considered judgment of the faculties and administrators of the educational institutions concerning the proper weighing of academic and non-academic information in admissions. Nor are the tactics employed by Nader and Nairn in advancing their point of view designed to lead to thoughtful debate of these issues.

Institutions using admissions tests believe that these tests have important benefits both to themselves and to the applicants. They do not view this as a narrow technical issue, though in large part statistical research confirms that the tests are useful. They are also well aware that the information and assessments of student characteristics they use in admissions are not infallible. Opportunities to contribute constructively to improvement of testing, assessment, and admissions are great. The Nairn report, which misrepresents many important facts and which offers no serious alternatives, contributes very little to such improvement.

Test Scores and Family Income

The Nairn report also asserts that ETS has attempted to suppress information on the relationship of test scores to students' family income, that the relationship of SAT scores to income is inordinately high, and that the tests preserve the social *status quo* by denying opportunity to students from poor and working class families. Each of these assertions is fallacious.

In several places, Nairn implies that ETS

is reluctant to publish information on the relationship of test scores to family income, and he alludes to "suppression of the score-income correlation."

In fact, the principal evidence reported by Nairn on SAT scores and family income is taken from a series of reports (**College Bound Seniors, 1973-74**) developed by ETS and published by the College Board. **College Bound Seniors** reports have been published since 1971-72, the year that the Student Descriptive Questionnaire was introduced. This questionnaire is the source of information on family income. More than 15,000 copies of this report are distributed each year to high schools, to colleges, to the press, and to the public generally.

Much of Nairn's discussion is clearly designed to leave the impression that test scores rank individuals according to their family incomes with few exceptions. In fact, the relationship is far more moderate than he suggests.

Table 2 shows average SAT scores and reported family incomes for the college bound seniors of 1973-74 for whom this information was available. The average income figures which are cited by Nairn show that there is, in fact, a relationship. Average family incomes are higher for higher-scoring students. Similarly, the average scores at the bottom of the table, ranging from 403 to 485, indicate that there is a relationship of scores to income.

The complete table shows that the relationship is far from perfect and that a ranking of students by SAT scores is not a ranking by family income.

Table 2
1973-74 College Bound Seniors Classified by SAT Average and Family Income*
Reported Family Income

SAT Average	$0-5,999	$6,000-11,999	$12,000-17,999	$18,000+	Average Income**
750-800	17	117	169	415	$24,124
700-749	239	1,172	1,752	3,252	$21,980
650-699	686	3,994	5,683	9,284	$21,292
600-649	1,626	9,352	12,187	17,992	$20,330
550-599	3,119	17,042	20,822	28,151	19,481
500-549	4,983	26,132	29,751	37,400	$18,824
450-499	6,663	33,209	35,193	41,412	$18,122
400-449	8,054	34,302	33,574	37,213	$17,387
350-399	8,973	29,762	25,724	26,175	$16,182
300-349	9,622	21,342	14,867	13,896	$14,355
250-299	7,980	10,286	5,240	4,212	$11,428
200-249	1,638	1,436	521	325	$ 8,639
Total Number	53,600	188,146	185,483	219,727	
Average SAT Score	403	447	469	485	

*The total number of students in this table (646,956) is very slightly smaller than the number (647,031) included in the analyses reported in **College Bound Seniors, 1973-74.** Students in this table must have had **both** SAT verbal and SAT mathematical scores and have reported family income on the Student Descriptive Questionnaire. Students with only one SAT score were included in **College Bound Seniors.**

From **College Bound Seniors, 1973-74.

Students from each income level obtain the full range of SAT scores. Many students from the top income group ($18,000 and over) earn low scores. For example, 8% scored below 350. Many students from the low-income group (less than $6,000) earn high scores—5% scored above 600.

What level of relationship might be expected between measures of students' educational development and their family incomes? It is well known that, in relation to students from low-income families, students from middle- and upper-class families usually have more highly educated parents, have home and community environments that provide more support for educational attainment, and attend better schools, to name only a few of the relevant factors. To deny that valid measures of educational attainment may be related to the economic circumstances of students' families is to ignore the realities of social and educational inequality.

Christopher Jencks (1972, p. 78), as cited by Nairn, estimates a correlation of about 0.35 between family economic status and scores on various elementary and secondary school standardized tests. A recent review by White (1976) found an average correlation of .25 between family socio-economic status and various indicators of educational achievement, based on 489 analyses and some 100 separate studies. The average of 41 correlations of socio-economic status and school grades was .24. When SES was defined solely in terms of family income, the average of 19 correlations between income and measures of educational achievement was .32.

The correlation found between SAT scores and income (of about .30) is quite consistent with more general research findings on the relationship of educational achievement to family circumstances and with the everyday experience of teachers in schools and colleges. The fact that we are not now all equal in educational development and achievement should not be obscured by heated charges of test bias and discrimination.

Tests and the Status Quo

A careful reading of Nairn's text and notes indicates that he does not in fact challenge the reality of the relationship of

students' family incomes and educational achievement. His fundamental thesis is that use of the tests, which help to disclose the effects of unequal resources and prior learning opportunities on the education of children of different classes, should be terminated or, at the least, modified.

Even if there were agreement that the best approach to expanding access to higher education is to eliminate evidence of unequal educational preparation, it is doubtful that the course of action advocated by Nairn would have the effects he predicts.

First, the relationship of SAT scores to family income is more modest than the statement implies and is not peculiar to the SAT. About 32% of the students with family incomes below $6,000 rank in the top half of the total group in terms of SAT scores (above about 450). Other similarly reliable and consistent measures of educational achievement would show a similar pattern. There is no evidence that use of test scores *per se* has a dramatic impact on opportunities for low-income students.

Second, admissions does not occur in the way Nairn suggests. Many colleges are not selective and admit nearly all applicants. Those colleges that are selective base admissions decisions on many different kinds of information, not test scores alone. In many cases, these colleges take into account the obstacles that disadvantaged students have overcome in reaching their present levels of achievement. Indeed, they seek out and provide financial aid and other kinds of assistance to such students.

Third, history indicates that selective admissions to higher education was far more a matter of class and economic status prior to the use of national admissions tests than it has been since. In the absence of a uniform and dependable indicator of students' abilities, admissions officers at selective institutions gave far more weight to grades and recommendations for students from a select group of well-known schools. The introduction of tests resulted in a substantial increase in opportunities for educational advancement of low-income students by providing a credible demonstration that many such students from schools without reputations for educational excellence could succeed in the demanding academic programs of the most selective institutions. Before rushing into radical surgery of the current system of admissions tests, we should carefully consider whether alternative systems would serve widely held social values as well.

As suggested above, admissions programs in selective colleges, and in many graduate and professional schools, are based on a balancing of values. These institutions place importance not only on achievement, accomplishment, demonstrated ability, and special talents, but also on more elusive personal qualities such as creativity and motivation. In their admissions procedures, many of these institutions also seek to redress effects of past inequality and to admit groups of students that are diverse in terms of geography, family economic background, race, and other characteristics. They do so not only to serve egalitarian principles, but also to accomplish their own educational objectives. When the **Bakke** case was before the Supreme Court, colleges and universities strongly defended their use of these kinds of criteria on social and educational grounds.

ETS and the educational associations that sponsor the admissions testing programs administered by ETS have also demonstrated in many ways their commitment to improving access of disadvantaged students to higher education. Very little attention is given in the Nairn report to the role of the College Board in advancing and developing through its College Scholarship Service (CSS) the concept of awarding financial aid based on need—historically one of the most important influences on expanded access to college for low-income students. Like CSS, a similar financial aid need analysis program at the graduate and professional school level, administered by ETS, is largely ignored in the Nairn report. Overlooked entirely are a host of talent search, guidance, scholarship, and demonstration projects designed to improve opportunities for disadvantaged and minority students—projects carried out by ETS on behalf of the test program sponsors or other organizations.

ETS is also committed in its research program to addressing root causes of differences in educational achievement for the poor and rich—differences that create the need for later programs of compensatory and affirmative action. Prominent on a long list of such research activities are a five-year study of effective compensatory reading programs in grades 2, 4, and 6 (sponsored by the U. S. Office of Education); a major longitudinal study of disadvantaged children and their first school experiences (sponsored by the Office of Child Development, Department of Health, Education, and Welfare); studies of exemplary school desegregation practices; and evaluation studies of the effects of educational programs on the skills and achievement of disadvantaged children, such as the **Sesame Street** evaluation (conducted for the Children's Television Workshop) and a number of local evaluations of compensatory education projects. This side of ETS is also given virtually no attention in Nairn's study.

Despite his slogan—that tests reflect "class in the guise of merit"—Nairn himself seems to recognize in his main argument the well-documented fact that **educational achievement** in general has a relationship to the economic background of students. Though this relationship is moderate, educational inequality is real. The failure of society to provide the best education possible to all its citizens has an impact on the capacity of individuals to lead satisfying and productive lives. This reality presents a challenge to society as a whole.

Nairn's proposal to eliminate evidence of inequality, before "constructing a society with a new definition of economic justice" is one that has its advocates. But, there are many who favor a more balanced response to this challenge—**first,** seeking through broadly based efforts to reduce real inequality in children's educational achievement; and **second,** recognizing within systems of advancement **both** the values of educational achievement and accomplishment **and** the need to expand access of all groups in the society to educational and occupational opportunities.

References

Anastasi, A., Meade, M. J., & Schneiders, A. A. **The validation of a biographical inventory as a predictor of college success.** College Entrance Examination Board Research Monograph No. 1. New York: College Entrance Examination Board, 1960.

Astin, A. W., King. M. R., & Richardson, G. T. **The American freshman: National norms for fall 1978.** Los Angeles: University of California, Los Angeles and the American Council on Education, 1978.

Baird, L. L. **Development of an inventory of documented accomplishments for graduate admissions.** GRE Board Research Report GREB No. 77-3R. Princeton, NJ: Educational Testing Service, 1979.

Davis, J. A. **Faculty perceptions of students. I. The development of the student rating form.** Research Bulletin 64-10. Princeton, NJ: Educational Testing Service, 1964.

Hilton, T. L., & Rhett, H. **Final report: The base-year survey of the national longitudinal study of the high school class of 1972. Appendix B, Part II.** Washington, DC: National Center for Educational Statistics, 1973.

Jencks, C. **Inequality, a reassessment of the effect of family and schooling in America.** New York: Harper & Row, 1972.

Nairn, A., & Associates. **The reign of ETS: The corporation that makes up minds.** Washington, DC, 1980.

Van Dusen, W. D., Nelson, J. E., Jacobson, E. C., & Ivens, S. H. **The College Board - AACRAO survey of undergraduate admissions policies, practices, and procedures: A special report on admissions requirements and test use.** New York: College Entrance Examination Board, 1979.

White, K. R. **The relationship between socio-economic status and academic achievement.** Doctoral Dissertation, University of Colorado, 1976.

Recommendations for the Improvement of Testing: Commentary on a Conference Report

by Frank W. Lanning

During the decade of the 70's, several major criticisms of standardized testing persistently challenged their use. The major criticisms were and still are:

1. That tests do not reflect the full range of student cultural backgrounds and thus lead to decisions that can be unfair to minority students.
2. That current standardized tests have only limited value for holding teaching, schools, and school systems accountable for the quality of education.
3. That tests exercise a limiting effect on classroom teaching.
4. That tests are too narrow in scope to provide for fair evaluation of new approaches to teaching.

In August 1978, responding to the need to address the criticisms, the National Institute of Education[1], HEW's educational research agency, sponsored a ten day conference to take a comprehensive look at testing and to make some recommendations. The conference's recommendations suggested some future guidelines for development and research in testing that could more effectively answer the major criticisms stated above.

I

To the criticism that tests do not reflect the full range of student cultural backgrounds the committee made these suggestions:

Better Fitting of Testing to the Cultural Background of Students

"Though there has been talk about the ideal of a culture-fair test since the beginning of testing, no test has yet been constructed which meets this ideal. Early expectations that perfect "fairness" or perfect standardization could be achieved by statistical means have proved unfounded. Under pressure from minority groups, test makers more recently have made some attempts to modify the language used in test instructions and test items so that it is more representative of vernacular language varieties. While these efforts are useful, it is by no means clear that a single test instrument can be equally representative of the language patterns of all major cultural groups in the population. Nor is it sufficient to modify surface linguistic features of tests. (p. 3)

"Recent advances in our knowledge of language interpretation and of cognitive processes reveal that culturally different experiences and background knowledge may affect test performance in complex and subtle ways. Take tests of reading skills for example. Investigators at the Center for the Study of Reading (Steffensen, Jogdeo, and Anderson, 1978) in well-controlled studies, showed the importance of the match between cultural background and reading passage content on reading speed, on reading comprehension, and on retention of information. Native American Indian children and majority culture children were given two stories to read about weddings. These stories had been carefully constructed to be comparable in vocabulary level and equivalent on measures commonly used as indices of reading difficulty, but one story dealt with wedding customs of Native Americans and the other with the "typical-American" wedding ceremony as depicted in magazines. Each group showed greater reading speed, superior recall, and better performance on comprehension questions for the story in its own tradition. This outcome seems commonsensical, but its implications for testing practices are profound. (pp. 3-4)

"While many items in achievement tests are designed to assess how familiar a student has become with material presented in the curriculum, other items use factual material or event descriptions in order to assess skills of comprehension, reasoning, memory, or problem-solving. Unless the material used for these purposes is equally familiar to all cultural groups, differences in performance are uninterpretable. The difficulties of achieving "equal familiarity" in this sense are so formidable as to make the ideal of culture-fair tests appear unrealizable and perhaps, unreasonable. (p. 4)

"An alternative strategy, and one adopted by many investigators specializing in comparative research, is to construct different forms of a "single test" — such that the difficulty level and knowledge domains assessed remain constant across forms, but the language and illustrative material are tailored to the specific background knowledge of particular groups. The aim here would be to devise "equivalent" culture-specific tests. This was, in fact, the strategy advocated by Binet, the inventor of the mental test. He rejected the notion that a uniform test could be used as a means of comparison of people from widely differing backgrounds and insisted that tests be appropriate to the background and everyday occupations of the individuals tested. Clearly the notion of perfect equivalence or comparability is also an "ideal" that can be more easily approximated for some types of tests than others. But participants in the conference were of the opinion that recent advance in knowledge and technique make this a fruitful strategy to pursue." (p. 4)

While conference participants were optimistic that recent advances in knowledge of language interpretation and of cognitive processes of culturally different experiences could eventually produce tests that would be cultural free, they, perhaps, failed to recognize the need for an extended period of research that would be free from political, government, and minority pressures that could bias and/or bring expedient closure on problems that will not be easily solved.

II

The committee's major recommendation of "better fitting of testing to education

Frank W. Lanning is a Professor of Education at Northern Illinois University.

objectives," was perceived to be basically related to three major criticisms of: (1) inadequacy for accountability, (2) negative or limited effect on classroom teaching, and (3) unsuitability to new strategies for teaching.

Better Fitting of Testing to Educational Objectives

"To understand how this lack of fit has come about, a look at the history of educational testing in this country is useful. The successful use of psychological and educational tests in World War I led to their wide adoption by school systems. Tests were used by the military services to select recruits for officer training and for training in various technical tasks. The method of sorting thus developed was then adapted to address the problem of sorting students in the civilian educational system. At that time, most students were not expected to finish high school and go on to college, and thus a major function of schools and colleges was to sort children and youths, encouraging only those who were most promising to go on. Much educational testing and testing theory developed in this context of sorting. In this theory, validity is measured by correlation of the test results with some other relative measure like grades in school and college. An item is judged by its ability to spread scores out for sorting purposes rather than for its relevance to what the school is seeking to help the student learn. (p. 12-13)

"In addition to these assumptions made to facilitate arranging students on a linear scale for sorting purposes, two other assumptions have tended to confuse and impede the improvement of educational testing. The first is the notion that the educational objectives of schools and colleges do not go beyond such simple skills as reading and computing and the recall of information in content areas. The second is the assumption that the attainment of important educational objectives can be adequately appraised by the use of paper-and-pencil tests alone. (p. 13)

"We have now moved into an era where we seek to help all students achieve their full educational potential. As discussed above, we are attempting to use tests for a range of purposes much broader than sorting of students. Work on the curriculum by teachers, scholars, and textbook authors in the last two decades has made explicit a wider range of educational objectives. The coming of low-cost information-handling technology makes it possible for us to escape the limitations of testing imposed by 50-year-old scoring technology based on multiple-choice paper-and-pencil tests." (p. 13)

This achieving of a better fit of testing to educational objectives could, according to the committee, be achieved through improved developments of three basic areas.

1. Criterion-referenced Testing

"A major response to the need for tests that serve purposes other than sorting has been the development of criterion-referenced testing. A criterion-referenced test determines what a student can or cannot do in a specified domain of educational objectives. Ideally, items are selected to give a proper representation of the domain. In contrast, traditional achievement tests, designed in the sorting tradition, eliminate test items that most students can answer, since these items do not produce the spread in scores desired for sorting. This latter practice tends to eliminate test items that represent what schools are trying hardest to teach and, as time goes by, may penalize better teaching by removing well-learned items in revised versions of the test. (These and other contrasts between criterion-referenced and ordinary achievement tests are discussed by Popham, 1978). Criterion-referenced tests can provide educators, parents, and others with a rather detailed picture of how well students, individually and in classes and schools, are learning in domains covered by the tests. Test makers may also develop data that allow comparison of performance in these domains among various categories of students and schools in different parts of the country. For example, the National Assessment of Educational Progress utilizes criterion-referenced tests and publishes national and regional performance data that schools can compare with their own. (pp. 13-14)

"Test makers are increasingly producing criterion-referenced tests. However, the theory and practice of constructing and interpreting the results of criterion-referenced tests need further development. Preparation of good criterion-referenced tests requires more careful analysis of the content domains being tested and preparation of more tests and test items than is the case with tests designed for sorting. Some tests presented as criterion-referenced are little more than re-worked versions of sorting tests, without the requisite coverage of content. Construction of criterion-referenced tests may still be strongly influenced by the traditional objective of spreading out the distribution of scores on a bell-shaped curve . . ." (p. 14)

Test makers are increasingly producing criterion-referenced tests. However, the theory and practice of constructing and interpreting the results of criterion-referenced tests need further development.

"Finally we note that although there are certain things that a school will wish every student to learn, we are also interested in encouraging students to develop special interests and capabilities of their own. Expecting every student to answer every question on a test is inconsistent with this goal and will tend to keep each student on a uniform path. A useful alternative (Zacharias, 1978) is to make items on a test 1/3 mandatory, 1/3 choosable from a longer list, and 1/3 designed by the student to show his or her grasp of an idea or skill. As children move towards the higher grades of school their interests and skills diverge—as they properly should in a country that lists 20,000 different titles in its dictionary of occupational skills. Something other than uniformatarian testing is needed to supervise and encourage the diverse growth of children's competences." (p. 15)

2. Information-handling Technology

The break-through here, as the committee saw it, has been the increasing availability of low cost information-handling technology. Some of the committee's major points for the discussion of three examples were as follows:

a. Computer-based Item Pools

"The task of providing teachers, schools, and school districts with tests closely matched to their specific educational objectives can be made manageable at reasonable cost through the new technology. Central computerized pools of test items of varying complexity could be created by educational centers and test publishers. Users could be given access to these pools either through direct communication between the central computer and their local computer or indirectly through local information storage devices such as magnetic or video discs. Items would be indexed in such a way that users could assemble them to form tests suited to their needs. The system could also provide additional information on the items, including national error rates and comments by other users or critics. It would be easy to design direct access systems in which additional items as well as comments and additional error data on existing items could be entered by any user, but an arrangement in which the center acted as an intermediary would probably be better. Similarly, many teachers might desire the help of a local expert in compiling tests. Tests could be printed out for paper-and-pencil administration, but an eventual further refinement could be administration and scoring of the test via a personal computer for each student." (pp. 15-16)

Computer-based testing can present problems realistically, can allow a student to proceed even when an arithmetical error is made, and can follow and evaluate the problem solving process.

b. Testing Problem Solving

"Technology now provides a means to improve instruction and testing in one of the ultimate aims of education—the ability to address more complex problems of the sort encountered in work and personal life. Such problems typically require us to bring to bear a variety of things we have learned in school or elsewhere; many are actually a series of problems, where each step depends on a previous step, and where various sequences of steps can be followed, some more efficient than others. Not only a student's answer, but also the efficiency of the solution strategy are of interest. Easily-graded standardized tests for these more complex problem solving procedures have been difficult or impossible to devise, so that a very important class of educational objectives has been left untested and thus undervalued. Information-handling technology now makes such testing possible at low cost. Computer-based testing can present problems realistically, can allow a student to proceed even when an arithmetical error is made, and can follow and evaluate the problem solving process." (p. 16)

c. Tailored Testing

"In a conventional test, the test taker works through a series of items in a fixed order, marking choice of response with a No. 2 pencil. One shortcoming of this procedure is that the test cannot measure accurately the abilities of test takers at the high and low ends of the score distributions for a heterogeneous group. In tailored testing, based on computer technology, the difficulty of the items is matched to the ability of the test taker. Multiple-choice questions appear on a display panel similar to a TV screen and the test taker indicates the answer on a typewriter-like keyboard. A correct answer is followed by a harder question, an incorrect answer by an easier one. No longer need the candidate waste time on items that are too hard or too easy, and responses to less appropriate items need no longer mar measurement based on more appropriate items. As the test proceeds, each response causes the computer to revise the estimate of the test-taker's ability. When the estimate reaches a specified level of reliability, the test ends (Urry, 1977). (p. 18)

"This approach is based on a "latent trait model" developed over 15 years ago (Lord, 1952) that makes it possible to give different forms of a test, all of different levels of difficulty, to subgroups of a population of candidates and to obtain comparable scores, as though all had taken the same test. Recently, investigators have been developing tailored testing procedures for use by the Civil Service Commission (Urry, 1977)." (p. 18)

3. Providing Better Information on Tests and Testing

The basic question that the conference wanted to discuss was the question of just what can a test do or not do. The committee believed that their first consideration was that of clarifying basic conceptions of testing. While the conference was not designed to resolve this problem they did state that:

"We believe, however, that the time is ripe for a new look at these basic questions in testing. This new look must involve, in addition to leaders in testing theory, persons from various branches of cognitive science as well as teachers and scholars concerned with the substance of education. The results of this effort should lead both to new research on testing and to informational materials for teachers, teacher's colleges, parent-teacher associations, and the general public, setting forth as clearly as possible what tests can and cannot do, including issues for continuing study and discussion." (p. 20)

The second consideration perceived important to this question was the appropriate use of tests in education. For this point the committee's statement was brief:

"Tests are more and more widely used in the educational system for accountability, selection, program evaluation, and instructional guidance. The use of tests has often become a routine bureaucratic practice to which little thought is given. Tests designed for one purpose are often used for other purposes to which they are not well suited, and critics argue that tests often have perverse effects even when used as intended. The conference believes a thorough program of studies of the use of tests in American education is needed to provide a basis for intelligent action. The studies should document the typical uses of tests by various educational agencies for various purposes and should investigate the effects of this testing on teaching, on the education of individual students, on educational innovations, and on budgetary and other decisions, and the appropriateness of these effects. These studies would make it possible to provide much firmer information on the suitability of various testing practices and how testing can be improved than is now possible." (pp. 20-21)

Private independent groups should be formed . . . to provide test users easy access to information . . . The model for such undertakings . . . something like *Consumer Reports* . . .

III

The third consideration for the conference concerned information on published tests. There were two recommendations that the conference felt were important concerning consumer information groups and full information disclosure of test items:

"**Consumer information groups.** Private independent groups should be formed, with private support, to provide test users easy access to information about the technical properties of published tests. A single group will not be enough: there must be several that evaluate tests from their points of view rather than seek a consensus or minimal evaluation. The groups would explain from their perspectives how the tests were developed, including the domain of knowledge and skills from which the items were selected, the formulation of items and the procedures used to reduce bias. They would describe the rationale for the content and format of the tests and the procedures for scoring.

"The model for such undertakings should not be the **Underwriter's Laboratory,** which approves the safety of an article, but something like **Consumer Reports,** which gives broadly accessible information about the factors of quality in an article and which then gives item-by-item information about how articles measure with regard to those factors. We are **not** recommending that tests be rank-ordered, or given ratings like "Acceptable", "Best Buy", etc. Currently, the Buros **Mental Measurement Yearbook** provides reviews of recently published tests, but these are given in a lengthy and detailed form that is probably not ideal for the unsophisticated consumer who wants to review what is available before choosing a test. We are not recommending that this activity be initiated by the federal government or a sponsored subsidiary such as an educational laboratory or center. Federal quality control over tests could evolve towards a national curriculum and in any event would be less desirable than several sources of test information developed through private initiative. (p. 21)

"**Fully informative disclosure of test items.** Students and parents affected by decisions based on tests should be able to see the tests—that is, the individual items

and the student's answers—on which these decisions were based. This would seem at once a fundamental human right and a necessity, since testing procedures are fallible. The research community and any concerned layperson should have easy access to test items and the grouped responses of test takers so as to form independent judgments about the quality of the items and the presence of possible bias, and freedom to publish the items on which their judgments are based. Arguments in a debate about an item or a test would seem to rest ultimately on appeal to items themselves, and without access to the items one cannot participate fully in the debate. (pp. 21-22)

"In considering the argument for secrecy, a distinction should be drawn between secrecy before administration of the test and secrecy afterward. Before administration, the topics to be covered may be revealed, but not the specific items and their answers. After administration, why should any matter be kept secret? Test publishers maintain a policy of secrecy after the test because the need constantly to redevelop old tests that had become public knowledge would increase the costs of test construction. One exception to the rule of secrecy is a federally-supported program, the National Assessment of Educational Progress, which releases 40 percent of its questions after each round of testing. Of course, most companies do release sample items which show some aspects of the questions asked. The computerized item pool described above requires no secrecy about items available for use on tests. We recommend also that the consumer information groups address themselves to the question of fully informative disclosure of test items." (p. 21)

IV

Combining Testing with Teaching

The interaction between teacher and student (the teaching process) is, perhaps, least served by tests now in use. Ideally, the use of tests for purposes outside the classroom such as accountability, selection, and evaluation needs to come out of classroom process, not be artificially imposed on this process. For this final major problem the conference identified four areas that could make significant contributions for more effective combining testing with teaching. Some of the conference's major recommendations follow:

1. Cognitive Science

"Cognitive science today is a rapidly-growing interdisciplinary movement including psychologists, computer scientists, philosophers, linguists, anthropologists and educators all converging on the analysis of human intellectual processes. An important component of this effort has been the detailed study of learning by students at all educational levels from preschool to college, coupled with a systematic effort to model partial and progressive states of knowledge using the computer. Any attempt to teach a subject involves some kind of theory or assumptions about what high-level performance in the area is like, as well as an explicit or implicit theory of the learning process. Contemporary cognitive science is conducting a much more searching appraisal of both performance and learning than has heretofore been possible. The work in cognitive science is far more theoretically compelling than the "learning theories" of twenty years ago and, at the same time, much more closely tied to direct studies of learning in school environments. Development of better systems of instructional guidance will be greatly aided by this knowledge of what is involved in the student's progressive encounters with subject matter. Presentations at the conference discussed efforts dealing with mathematics, reading, and writing, some of which will be described here. (pp. 22-23)

"One of the attributes of a good teacher is the ability to diagnose underlying misconceptions from a student's answers to a set of problems. The BUGGY computer program, developed by John Seely Brown and Richard R. Burton (1978), can undertake this task for certain aspects of mathematics instruction. The computer program includes "correct models", which represent the various ways to obtain correct answers, and "diagnostic models", which represent various ways that students typically obtain wrong answers. Like a good teacher, BUGGY is not limited in instructional guidance simply to indicating which answers are correct and which wrong, but can also indicate which of many misconceptions a student may harbor: the "bugs" in his procedures. There are many possible bugs in children's arithmetic (e.g., always subtracting the smaller digit from the larger: 1928–573=1455). Brown and Burton are devising a computer program allowing diagnosis of more than one bug, a task generally beyond the capability of a teacher. (p. 23)

"It is tempting to assume that students make mistakes because they do not follow procedures very well: that the primary cause of error is simply inability to carry through a sequence of steps properly. But good teachers operate on the assumption, and the BUGGY program begins to demonstrate, as Brown and Burton note, that students are remarkably competent followers of procedures; the difficulty is that they often follow the wrong procedure. That is why cognitive science can aid good teaching by investigating not only the student's answer to a set of problems, but also the process by which he or she obtained those answers, so that the correct processes can be learned." (p. 23)

2. Interactive Teaching-Testing and Technology

"Generally, human beings learn best when there is rich and immediate response to the learner's initiatives and when there is the possibility of branching, exploratory interplay between the learner and the teacher and testing is part of the interplay. Most schools are not able to provide that kind of learning environment because it would require something close to a one-teacher-one-student relationship. The rapidly increasing capacity and decreasing cost of electronic information-handling technology make interactive teaching-testing arrangements both feasible and economically attractive. (p. 24)

"Conferees proposed exploration of a new vision of a learning and testing environment extending what has been just described. In this vision of the future, school tests as we know them would cease to exist. The intrusive, specialized, institutionalized activity called testing would be absorbed into a new kind of learning and testing environment. Computers could accept inputs from students and teachers on an almost continual basis, extracted from the rich tapestry of ongoing learning activities. Instructional systems would accumulate an educational portfolio for each student, including a wide range of interrelated performance and situational descriptions. One would be as unlikely to cease all instructional activities in order to test a student as one is to stop conversing with a child in order to test his or her linguistic competence. Instead, testing would be a particular aggregation and analysis procedure applied to a continuously collected data base. Some of these aggregation would have an immediate impact on ongoing learning activities; other would be remote from the moment of data collection." (p. 25)

The recommendations also suggest that interactive teaching-testing materials for schools could also be developed without a computer. This basically can be done with the development of a variety of self-checking tests.

3. Subject Matter

"To construct good tests to assess what a student knows and can do in a domain of knowledge requires not only skilled test developers, but also outstanding teachers and scholars experienced in the content domain as part of the testmaking team. Persons with such background who look through collections of published tests in

their area typically find in many of the tests numerous items that they regard as inappropriate exemplars of the knowledge area, as seriously ambiguous in working, and too often as simply wrong. This leads them to conclude that knowledgeable persons were not involved in creation of the tests in question. Involvement of knowledgeable teachers and scholars in testmaking, important for any test, becomes absolutely essential in the development of testing-teaching systems and tests of problem-solving processes. The conference heard papers from teachers and scholars in the sciences, mathematics, and reading and writing which discussed educational objectives and intellectual processes in their areas with a richness and depth that only persons with their knowledge can bring to the testmaking process." (p. 26)

4. The Natural Classroom Situation

"The teaching-testing systems envisioned . . . are intended as aids to the teacher, rather than as substitutes for the teacher. Systems developed on this basis must take account of the natural classroom situation or be in peril of being rejected or poorly used. Developers should make full use of what is already known about the way instruction and instructional guidance take place in typical classrooms at the grade level and in the subject area in question and about the successes and failures of other newly-introduced systems. However, the new system envisioned is significantly different from what has gone before, and its reception by teachers and students is not predictable in detail. Research on these matters must be built into development projects and be used to improve the system. (p. 26)

"The number of investigators interested in direct, intensive observation of classrooms is growing. They include anthropologists, psychologists, sociologists, and teachers with a talent for stepping back from their experiences and describing them. Conference participants working in this area provided several examples of how tests could be made more useful in the classroom. Test results closely related to the day's activities rather than to larger units are likely to be most useful to both the teacher and the student, and feedback on what probably went wrong would be even more useful. However, the need is not uniform across students. Teachers often believe they have enough information about many students in the classroom. For some students the teacher may feel the need for a great deal more information: the teacher doesn't understand the student, has "tried everything and doesn't know what to do." Testing procedures providing insight into such students' intellectual functioning could be very helpful to teachers. Both examples illustrate the more general principle that teachers, like other people, prefer arrangements that adapt to their needs rather then requiring them to adapt themselves to a rigid system. We do not wish to imply, however, that new systems should respond only to needs currently perceived by teachers. The most useful innovations may be those undreamed of prior to their invention, but they too must be tried in real situations and modified as necessary." (p. 27)

In Conclusion

This conference was significant. Its recommendations and papers are important to test development of the 1980's. The conference, if it did not resolve the major tests criticisms of evaluation of new approaches to teaching, limiting effect on classroom teaching, accountability for quality of teaching, and failure to reflect the full range of student cultural backgrounds, did, at least, identify the ways and means of dealing with these major criticisms. Continued support for research on classroom processes and human cognition and for development of new technologically-based testing was a fundamental conference conclusion. The full report of this conference contains a great deal of information that could be useful for the consumer, teacher, and developer of tests.

References

Brown, J. S. & Brown, R. R. Diagnostic models for procedural bugs in basic mathematical skills. **Cognitive Science,** 2: 155-92. (Reproduced in Part Two of this volume), 1978.

Buros, O. K. **The Eighth Mental Measurements Yearbook,** 2 vols., Edison, N.J.: Gryphon Press, 1978.

Burton, R. R. & Brown, J. S. An investigation of computer coaching for informal learning activities. **International Journal of Man-Machine Studies,** 11:5-24, 1979.

House, E. R., Glass, G. V., McLean, L., & Walker, D. F. No simple answer: Critique of the "follow-through" evaluation. **Harvard Educational Review,** 48:128-160, 1978.

Lord, F. M. A theory of test scores. **Psychometric Monograph,** No. 7.

National Council of Supervisors of Mathematics, Position paper on basic mathematical skills. **The Mathematics Teacher,** 71:147-152, 1978.

Popham, W. J. The case for criterion-referenced measurements. **Educational Researcher,** 7:6-10, 1978.

Porter, A. C., Schmidt, W. H., Floden, R. E., & Freeman, D. J. Practical significance in program evaluation. **American Educational Research Journal,** 15:529-539, 1978.

Steffensen, M. S., Jogdeo, C., & Anderson, R. D. A cross-cultural perspective on reading comprehension. Technical Report No. 97. Champaign, IL: Center for the Study of Reading, 1978.

Throndike, E. L., et. al. **The Measurement of Intelligence.** New York: Teachers College Bureau of Publications, 1927.

Urry, V. W. Tailored testing: A successful application of latent trait theory. **Journal of Educational Measurement,** 14:181-196, 1977.

Zacharias, J. R. **A discussion of tests that concern the sciences and the technologies.** Testing, Teaching and Learning: Report of a Conference on Research on Testing, Aug. 1978, National Institute of Education, Oct., 1979, pp. 54-59.

Footnote

[1]This commentary was based on "Chairmen's Report," in **Testing, Teaching and Learning: Report on a Conference on Research on Testing,** August 17-26, 1978, Ralph W. Tyler and Sheldon H. White Chairmen. Sponsored by the U.S. Department of Health, Education, and Welfare/National Institute of Education. Published Oct., 1979, pp. 1-28.

MEDIA CORNER

by Peter C. West

Roger Lennon of the Harcourt Brace Jovanovich Test Department has estimated there are approximately 200,000,000 standardized tests administered annually in the schools throughout the United States, and that this utilization will undoubtedly increase in the years to come. He further states "the sheer volume of standardized testing calls for the most thoughtful attention to selection and provision of the best testing materials."

In addition to the actual tests, many materials have become available in recent years for familiarizing oneself with standardized testing and related standardized testing procedures. The following is but a partial listing of the many excellent media resources available for preparing educators to make a thoughtful decision in the selection of the best testing materials.

16mm Films

The I.Q. Myth (Carousel Films)
Purchase $625.00

This film examines how the I.Q. Test has been used, misused and abused through the years. The film focuses on the question of how much importance, if any, should be placed on the results of a single test and points out that such exams have been used to categorize entire cultures and races as mentally inferior. 51 minutes.

Peter C. West is Assistant Director, Learning Center, Northern Illinois University

Sound Filmstrips Sound Slides

Intelligence Testing
Multi-Media Productions, $15.00

This program proposes to take a brief look at the history and nature of IQ tests to show the problems inherent in their widespread usage.

Minimum Standards & Competency Based Education
Phi Delta Kappa, $20.00

This sound/slide program is a status report on minimum competency testing. It reviews the 36 state mandates for competency testing and analyzes and compares them.

Testing, Testing, Testing: How to Get Better Grades
Guidance Associates, $69.50

The sound filmstrip program explains two types of tests - curriculum based and standardized tests. It explains how standardized tests are used to measure interests, aptitudes and abilities.

Audio Tapes

How Tests Fail; What Test Users Need to Know About the Limitations of Standardized Tests
University of California, $90.00

An abundance of basic, nontechnical information on the nature of standardized intelligence and achievement tests and their limitations.

Exercises in Classroom Measurement
Charles E. Merrill, $125.00

Set of 11 cassettes describing 25 different exercises. Among the exercises are "Use of Buros's Mental Measurements Yearbook," "Use of Test Manuals and Norms," and other exercises for standardized testing.

Addresses:

Carousel Films
1501 Broadway
New York, New York 10036

Charles E. Merrill
1300 Alum Creek Drive
Columbus, OH 43216

Guidance Associates
Box 300
White Plains, New York 10602

Multi-Media Productions
P.O. Box 5097
Stanford, CA 94305

Phi Delta Kappa
8th and Union Ave.
Box 789
Bloomington, IN 47402

University of California
Extension Media Center
Berkeley, CA 94720

Messages & Markets

Classified Rates: up to 50 words, $8.00; 51-100 words $15.00; Address:
Business Manager
Thresholds In Education
P. O. Box 771
DeKalb, IL 60115

Perception

Interested in Perception? Perceptual limitations? Learning Problems? So is the **Association for the Study of Perception.** Journal published. Conferences held. For Information write to: Association for the Study of Perception, Box 744, DeKalb, IL 60115.

We invite you to join . . .

Mind/Body Education
A Special Interest Group of
The American Educational Research Association

Recent advances in the study of consciousness and the potential of the human organism suggest that vast new areas of human behavior are available for systematic research and investigation. These advances outline new views of learning, knowing, teaching and therapy and raise the distinct possibility that current views of education are inadequate. Several sources of these advances are biofeedback, left-right hemisphere studies, state-specific studies (e.g., dreams, several kinds of meditation, yoga, hypnosis, suggestion, imagery, and mind drugs), and latent abilities. The purpose of the SIG: MIND/BODY EDUCATION is to undertake and promote research into these and related areas to determine their educational significance and implications.

Both members and nonmembers of AERA are welcome in this SIG. We hope that teachers, parents, counselors, and interested citizens will join to help us bring these approaches to education at all levels. Dues are $4.00 a year, May 1 - April 30th.

Secretary-Treasurer

Marianne W. DeVoe
1709 Split Ridge Road
Knoxville, TN 37919

Tulane Launches Project Talent

New Orleans . . . Tulane University has established Project Talent: A Program for Gifted Younger Students.

The establishment of Project Talent was announced by Tulane University president Sheldon Hackney. He pointed out that Project Talent is one of the few programs in the country—and the only program in the Sun Belt—that enables superior students of younger than normal college age to take regular university courses for full credit.

"While many universities have opened enrollment to older students, we are pleased that Tulane has created a program that meets the special needs of younger students," President Hackney said. "Moreover, the presence of younger students on campus adds an exciting new dimension to the entire university."

Project Talent begins with the university's summer 1980 session, when 25-30 exceptionally bright, motivated and mature youngsters will join Tulane undergraduates as full or part-time students, on a residential or non-residential basis.

"By offering this opportunity to gifted and talented young people, Project Talent challenges them academically and thereby helps to alleviate the tedium and frustration they may find in some of their regular school work," explains Fred Zuker, director of Tulane Admissions and coordinator of the project.

"In addition, Project Talent students earn college credit. By completing their formal education early, they are able to make their distinctive contributions to the world sooner," he added.

Golden, CO—The Second Edition of the **National Solar Energy Education Directory** has been published by the Department of Energy's (DOE) Solar Energy Research Institute (SERI). It is priced at $5.50.

The 200-page Directory is a comprehensive up-to-date listing of 1,760 solar-related courses and 240 programs and curricula offered at over 750 post-secondary educational institutions nationwide. SERI specialists anticipate that the publication

will be of particular value to science teachers and guidance counselors in assisting students selecting higher education as preparation for solar-related careers.

For a copy of this Directory, write:

Superintendent of Documents
U.S. Government Printing Office
Washington, D.C. 20402

Summer Session Announcement

Northern Illinois University's summer session will be conducted from June 16 through August 8, 1980. Information concerning admission, courses, and special programs is available through the College of Education Advisement Office, Gabel 153, Northern Illinois University, DeKalb, Illinois, 60115 (815-753-1541).

Holding a conference or workshop next summer or fall?

Send your announcements to Business Manager,

Thresholds In Education
P.O. Box 771
DeKalb, IL 60115

The **Pourchot Mechanical Manipulation Test** is designed to test bi-lateral hand-tool dexterity. The task of relocating pipes and bolts in holes provided is accomplished through the use of common tools: i.e. pipe wrenches, a screw driver, and a small wrench with multiple openings. Coordination of both hands in the use of the tools is required to complete the task efficiently. The test and the accompanying tools are of substantial size and weight so as to give reality to the manipulative tasks. The skills required are related to those needed for assembly and repair jobs in automotive, farm, home, plumbing and various industrial jobs. Published by:

Friendship Supplies
P.O. Box 82
DeKalb, IL 60115

The Prairie Heritage

Read about it in the
GURLER CHRONICLES

No. 1 **Natural Prairie in the DeKalb Area, 1830 to the Present,** by Don Murray. What the first settlers saw and plowed under—and present efforts to preserve and restore the prairie.

No. 2 **Taming the Wild Prairie, Northern Illinois, 1830-1900,** by Ruth Shonle Cavan. The social development from farm to small industrial city.

No. 3 **Music of a Young Illinois, 1830-1900,** by Mary Ellen Pourchot. Ballads, Civil War songs; vaudeville, the town band, and concerts; melodeons and pianos, and finally the phonograph.

Each CHRONICLE: $1.00, plus 30¢ mailing cost. Gurler Heritage Association, 205 Pine Street, DeKalb, IL 60115.

Book Review

Introduction to Research in Education. Donald Ary, Northern Illinois University; Lucy Cheser Jacobs, Indiana University; Asghar Razavieh, Pahlavi University, Shiraz, Iran. Second Edition, Holt Rinehart and Winston, 1979. Reviewed by Linda J. Burke, Instructor, Department of Learning and Development, Northern Illinois University.

Introduction to Research in Education, second edition, by Ary, Jacobs, and Razavieh, is designed as a basic research text for undergraduates and masters level students. The introduction of the text sets out two goals for students: "1) to understand and evaluate the research of others, and 2) to plan and conduct their own research with a minimum of assistance" (Ary, Jacobs and Razavieh, 1979, p. vii.) This reviewer feels that the text accomplishes both goals, but probably succeeds better at the second goal than at the first.

The chapters cover the broad range of topics needed to meet these dual purposes, but the chapters are organized into parts in the order in which the research process occurs. The student is thus guided step-by-step through the stages of problem formulation, design selection, etc., which the typical beginning researcher follows. Part 1 lays the foundation by discussing the scientific method in contrast to other ways of knowing. Part 2 begins the "meat" of the book: defining the problem, searching the literature, and stating the hypothesis in operational form. The student has the terminology defined clearly and is given realistic examples to guide thinking. Both descriptive and inferential statistical analyses are covered in Part 3, if the instructor wishes to cover these topics in the course. The section on sampling at the beginning of Chapter 6 is a lucid discussion which provides necessary background for the rationale of hypothesis testing. Measurement concerns related to use of definitions are outlined in Part 4. Understanding the concepts of reliability and validity is a prerequisite for the selection of measurement instruments appropriate to the research study. The application of principles of research design can be a difficult subject for the beginning researcher. Pros and cons of the various types (experimental, *ex post facto*, historical, descriptive) are given and illustrated. Examples of common pitfalls, such as threats to internal validity, should help the student avoid these in research. The coverage of these topics is not as extensive as in some competing texts (Borg and Gall, 1979, is an example). Instead, the student is presented the basics in a solid way, without getting bogged down in minutiae. Part 6 ends the text appropriately with data analysis techniques and preparation of the report.

The arrangement of the text helps the student who is expected to formulate a research proposal, possibly for the first time in her academic career. No background information is assumed, which is a real advantage of this text. Obviously, the principles by which good research is designed and by which it is criticized are the same. But the Ary, Jacobs, Razavieh, text does not give any specific examples of a critique of an article, nor does it give guidelines as to the compilation of the literature review. An instructor stressing these topics in a course would need to provide supplementation here. The text also does not deal (appropriately enough) with some peripheral topics such as curriculum evaluation and participant observation research. The text is an introductory one, covering every **major** component of educational research. Observational techniques, sampling, and deductive reasoning are just a few of the topics touched on. The stress in the text is on the basics, the essentials, and the student is spared wading through a mass of material.

Several changes which have been made from the first edition have improved the usefulness of the text. Study exercises (and the answers) have been added at the end of each chapter. None of the other texts which were examined (see references) has done as thorough a job on the use of the ERIC resources and computer searching—an important source of information for the educational researcher. Correlational analysis is common in educational research, and the authors have attempted to lay a strong conceptual foundation in Chapter 5. Most noteworthy is the expansion of Chapter 10 on *ex post facto* research. Most other texts relegate this topic to minor status, lumping it together with either experimental, or descriptive research. Because many variables used in educational studies cannot be manipulated, *ex post facto* research is important. It is often undertaken, however, without an awareness of the difficulties inherent in proper interpretation of the results. Problems in the determination of causality are illustrated by specific examples, as are ways of achieving partial control.

In the last section, two topics have been given more extensive coverage than in the first edition. The discussion of research ethics goes beyond the basic caveats of protection of rights of subjects, and gives current DHEW guidelines. The role of institutional review boards is also mentioned. Computers are now the name of the game in data analysis, and the beginning student is familiarized with basic terms and procedures. Suggestions as to use of computers are given, as well as warnings as to what the computer will **not** do for the researcher.

Evaluation of the worth of a text to the student cannot rest only on the up-to-dateness and extent of coverage of content. Style, clarity, and readability are important considerations. Of course, these judgments are more subjective and less easily supported. The Ary, Jacobs, Razavieh text is to be applauded for its efforts to eliminate sexist language while retaining clarity of statement.

Despite the three authors, style is consistent throughout, and it would be difficult to select which chapters were written by each. Readability is enhanced by the limited use of commas and extended clausal structure. Examples are used judiciously. They are present when needed, and appear as an integral part of the text. The second edition does have more examples than does the first, but they are not overdone. The selection from "Zen and the Art of Motorcycle Maintenance" relating to the use of the scientific method is a relevant addition of this type.

The reviewer found the first edition a first-rate text, and popularity of the book attests to the fact that others shared that view. The second edition now awaits the assessment of current education students. I think they will come to appreciate it in the same way the first edition was accepted.

References

Ary, D., Jacobs, L., & Razavieh, A. **Introduction to research in education.** New York: Holt, Rinehart and Winston, 1979.

Best, J. W. **Research in education** (third edition). Englewood Cliffs, New Jersey: Prentice-Hall, 1977.

Borg, W. and Gall, M. **Educational research.** New York: Longman, 1979.

Dyer, J. **Understanding and evaluating educational research.** Reading, MA: Addison-Wesley Publishing Company, 1979.

Sax, G. **Foundations of educational research.** Englewood Cliffs, NJ: Prentice-Hall, 1979.

Tuckman, B. **Conducting educational research.** New York: Harcourt, Brace & Jovanovitch, Inc., 1978.

What's Happening in the State of Illinois in Standardized Testing — A Study

The Status of Standardized Testing in Illinois

by Carol Mardell-Czudnowski & Norman Stenzel

In Illinois, what tests are most widely used? To whom are they given? Why are they given? How much does testing cost?

Formal tests constitute an integral part of the American educational process. They have permeated every grade level and subject area, from preschool through the university, from reading to adaptive behavior. The purpose of this article is to describe and explain the status of standardized testing in Illinois public schools: what tests are most widely used; to whom, from kindergarten through senior high (grade 12), are they given; why (the perceived purposes); and the cost of this practice.

In order to comply with a legislative directive to determine the status of testing in Illinois, the Program Assessment and Evaluation Section of the Department of Planning, Research and Evaluation of the State Board of Education devised, sent out, and collected survey information about testing in Illinois school districts. As of December 1979, 973 school districts (over 95% of the districts in the state) responded to the "Status of Testing Survey" distributed in September, 1979. Thus, this report can be considered an accurate statewide description of educational testing.

The survey requested information about various aspects of testing programs in the public schools. This included tests used, grade levels, purpose of testing, status and activities related to minimum competency testing, and cost of testing. It should be kept in mind that this report focuses on testing on a district-wide basis; additional local testing less extensive than district-wide in scope also takes place, but is not reported here. For instance, tests used in special education are administered on an individualized basis. These have been surveyed separately. (Mardell-Czudnowski, 1980).

Carol Mardell-Czudnowski is an Associate Professor of Education at Northern Illinois University.
Norman Stenzel is with the Illinois State Board of Education

Most Widely Used Tests

District-wide testing must be divided into two categories: locally developed and commercially developed. Some sort of locally developed district-wide testing currently occurs in all of the districts in the state and deals most commonly with mathematics. Reading or reading readiness ranks second in frequency of use. Other areas of testing included writing, science, and citizenship. The use of some commercially developed test batteries takes place in 89% of the school districts of the state. The most widely used standardized test batteries are: SRA Achievement Series; Stanford Achievement Test Series; and Iowa Test of Basic Skills (in that order). These are used at the elementary level. Widely used tests at the high school level are the National Educational Development Test and the Iowa Test of Educational Development.

Commercially developed reading tests are used in two-thirds of the school districts of the state. The most popular tests at the elementary level are the Gates-McGinitie Reading Test and the Stanford Diagnostic Reading Test. At the high school level, the Gates-McGinitie is the most commonly used instrument.

Approximately one-fifth of the school districts in the state use commercially developed mathematics tests. The two tests more frequently used are the Key Math Diagnostic Arithmetic Test and the Stanford Diagnostic Mathematics Test.

Grade Levels of Children Tested

Locally developed testing most frequently takes place in grades three through eight. Commercially developed testing is limited by the grade levels upon which the tests were normed. Over 20% of the districts, for instance, use the Metropolitan Readiness Test in kindergarten. No other commercially developed test has wide usage at this grade level.

Commercially developed standardized test batteries are widely used between first and eighth grades. Considerably less testing of this type is done at the high school level. This trend holds true for commercially developed reading and mathematics tests as well.

Purposes of Testing

The purpose of locally developed district-wide testing most frequently is for diagnosis and remediation. The next most frequent use is for program or curriculum evaluation. The third ranking use of locally developed district-wide tests is for placement of students.

The purposes reported for commercially developed standardized test batteries were the same as for locally developed tests. Placement, however, became more important than program or curriculum evaluation when reporting the purpose for using commercially developed reading or mathematics tests.

Minimum competency testing is another purpose for testing. Such testing is defined as "the use of a uniform measurement instrument(s) which is constructed to measure the number of competencies or skills which students have acquired which are deemed to be minimally acceptable within the school district" (Stenzel, et al., 1979). This can be accomplished by use of a locally developed test or a commercially developed test.

Results indicate that prior to 1979, 8.7% of Illinois school districts had some form of minimum competency testing. By December of 1979 that had increased to 9.6%. At the end of the 1979-80 academic year, the number is expected to increase to 16% of the school districts in the state. At this time, with developmental efforts and a phase-in period considered, the reports from districts indicate that nearly 30% will have a minimum competency test by the end of 1982.

By reviewing the purpose of testing information on the survey, if one considers the purpose of minimum competency testing to be graduation, promotion, and diagnosis-remediation, it can be estimated that as many as 50% of the districts in the state use

locally developed district-wide testing in a manner similar to minimum competency testing. A similar review of the use of commercially developed tests suggests an estimate that over 80% of the school districts in the state currently use commerically developed tests for purposes similar to those of minimum competency testing.

A few districts (3.2%) report that they have determined not to implement a minimum competency test. Other districts have not yet made a decision; however, nearly all of the districts (94.6%) had discussed minimum competency testing by December, 1979; and three-fourths (75.4%) had Board of Education discussions. Other activities related to the consideration of minimum competency testing were also reported in the survey. Districts (60.1%) have had staff attend workshops, and some (41.8%) have reviewed commercial tests.

Cost of Testing

A set of six questions on the survey dealt with the cost of testing in districts during the 1978-79 school year. Based on that information, on a per pupil basis, the cost of commercially developed testing and locally developed testing was similar during 1978-79. The cost per student for commercially developed testing for those districts having an expenditure was $2.01, and the cost per student for locally developed testing for those districts with an expenditure was $1.93. Districts may not have an annual expenditure for purchase or development of tests. Tests purchased from commercial sources, for example, can be reused and scored locally; and once local development has taken place, there are few costs which might appear in school ledgers after that. Obviously, indirect costs such as space, lighting, and personnel time for administration and/or scoring are not included in these figures.

Total reported expenditure for district-wide testing during 1978-79 was $2,800,143.20. Commercially developed testing alone cost $2,113,931.80. Locally developed testing cost $686,211.40. The total cost of test development from initial activities to the present for districts with locally developed district-wide tests was $1,594,618.40.

. . . although there are a large number of tests from which to choose . . . The 39 tests most frequently used are published by nine firms.

It can be readily seen from these figures that a great deal of money is expended annually on commercially developed tests. It is interesting to note that although there are a large number of tests from which to choose, there are a small number of publishing firms. The 39 tests most frequently used are published by nine firms. The ten most widely used tests mentioned in this article are published by only four firms.

Discussion

Currently, standardized testing on a district-wide basis, whether locally or commercially developed, is a very popular practice in the state of Illinois. Millions of dollars are expended annually to obtain information for purposes of diagnosis-remediation, program or curriculum evaluation, placement of students and minimum competencies. Further study should be made of the appropriate use of a number of tests mentioned as being used on a district-wide basis. Such tests as the Peabody Individual Achievement Test, Slosson Oral Reading Test, Woodcock Reading Mastery Test, Wide Range Achievement Test, and Key Math Diagnostic Arithmetic Test are all basically individually administered tests. Administering such tests to every student on a district-wide basis would be a questionable practice, both in terms of testing efficacy and cost-effectiveness. In addition, some school districts report using commercial tests at inappropriate grade levels or for inappropriate purposes.

The Illinois State Board of Education serves as a resource to local school districts. Unfortunately, many districts do not take advantage of the available services to solve testing issues.

Further study also needs to be directed towards the development of local achievement and minimum competency tests. Such tests should be criterion-referenced rather than norm-referenced, although local norms can also be easily established. These tests obviously need content validity, but they also need to be evaluated according to other basic test construction principles to establish reliability and appropriate cut-off points for sensitivity and specificity. Typically, locally developed tests often lack these important test construction characteristics.

The Illinois State Board of Education serves as a resource to local school districts. Unfortunately, many districts do not take advantage of the available services to solve testing issues. For instance, only 13.1% have reviewed the Performance Indicators for Competency Assessment (PICA) Notebook developed by the state office and only 7% have applied for grants from the Illinois Centers for Educational Improvement. Thus, a better working relationship seems necessary in order to resolve some of the testing issues which still are apparent at the district level.

References

1. Mardell-Czudnowski, C. The four Ws of current testing practices: Who; what; why; and to whom -an exploratory survey. **Learning Disabilities Quarterly**, 1980, 3, 1, 73-83.

2. Stenzel, N., et. al. **Status of testing survey**. Springfield, IL: Illinois State Board of Education, 1979.

Teacher Supply and Job Demand for the Early 1980's

by Leonard L. Pourchot

How many students will be in school in the 1980's?

How many teachers will be needed in the years ahead?

In what subject fields will demand be greatest?

Where in the United States will jobs be available?

These questions perennially interest educators, teachers and students. The Association for Schools, Colleges and University Staffing (ASCUS) made a study in 1976 culminating in a report, "A Current Assessment of Teacher Supply and Demand in the United States." A follow-up study by ASCUS is reported in a research report "Teacher Supply/Demand in the United States 1980." The statistics which follow are based largely upon the latter report.

How many students can be expected?

The Division of Statistics, Natality Statistics Branch, Department of Health, Education and Welfare, reported a decline in live births in the United States from 1960 to 1973. A low birthrate persisted through 1976, but a modest increase occurred in 1977 and 1978 which appeared to continue through 1979. Assuming that the modest increase in births continues, a slight increase in the demand for elementary teachers may be seen in the mid-1980's, but these slight increases would not be seen in high schools until after 1990.

How many teachers will be needed in the next several years?

During the 1970's, teacher production shrank to half the size of the peak year of 1971-2. Production of both elementary and secondary teachers has been cut back, but secondary teaching has experienced the more severe reduction of numbers. Apparently supply/demand fluctuates by geographical area and by subject fields. During the 1970's, severe cutbacks in teacher preparation occurred throughout the United States, but from 1978-79 to 1979-80 it appears that the situation has largely stabilized. In fact, in '79-'80 the number of total teachers prepared has increased somewhat in several states. In view of the birthrate's decline, it seems unlikely that large numbers of teachers will, or should, be prepared beyond replacement needs (See Table A) (1).

In which fields are the greatest needs?

According to ASCUS, placement directors around the country perceive teacher shortages in the following subject areas:

agriculture
business education
chemistry
distributive education
earth science
general science
industrial arts
mathematics
physics
reading
school psychologist
special education
speech correction

The placement directors perceive surpluses in:

art
elementary education
health
physical education - men
physical education - women
social studies

Other subject areas show more of a balance between supply and demand.

Where in the United States are teaching jobs available?

Reports from placement directors showed widely scattered reactions, but some generalizations may be drawn. Elementary and junior high supply and demand seemed balanced except for shortages reported in scattered areas including Colorado and Wyoming. Mathematics, natural and physical sciences, trade, industrial and vocational-technical teachers, and special education teachers were reported in short supply in all geographic areas. The same was true for industrial arts, reading, speech correction and school psychology.

Unfortunately, the teaching fields in which there are surplus teachers or a balanced supply seemed to exhibit this abundance in almost all geographic areas. Only two states, for example, reported shortages of social studies teachers: North and South Dakota. Several states reported shortages of English/language arts teachers: Colorado, Illinois, Indiana, Iowa, Kansas, Missouri, North Dakota, Ohio, South Dakota and Virginia.

One institution which prepares teachers had the following numbers of graduates earning degrees in education for 1978-79 in the various subject fields (2). See Table B.

Table B shows that of 691 teacher candidates earning degrees in education in 1978-79 from one institution, 345, or about 50% were preparing in fields judged to have surpluses in most areas of the country. Fields such as chemistry and other sciences and mathematics in which widespread shortages were reported had comparatively few teaching candidates. If the experience of the one institution is indicative of what is happening in similar institutions, both the teacher shortages and the teacher surpluses may persist into the future.

References

1. ASCUS Research Report. **Teacher Supply/Demand in the United States 1980.** March, 1980.
2. Career Planning and Placement Center. Northern Illinois University. "Placement Data for December 1978—May/August 1979 Undergraduates Earning Degrees in Education."

TABLE A[(1)]

A COMPARISON OF PERCENTAGE CHANGE IN NUMBER OF TOTAL NEW TEACHERS PREPARED BY STATE FOR SELECT YEARS

	1971-72 vs 75-76	1975-76 vs 78-79	1971-72 vs 78-79	1978-79 vs 79-80
Alabama	-25.62%	-03.65%	-28.34%	-06.07%
California	-13.85%	-35.86%	-44.74%	-01.17%
Colorado	-46.80%	-27.33%	-61.34%	-06.42%
Florida	-13.56%	-02.36%	-14.53%	-09.82%
Georgia	-38.46%	-45.83%	-66.66%	+27.77%
Idaho	-49.53%	-12.96%	-56.07%	00.00%
Illinois	-19.97%	-35.12%	-48.08%	-07.45%
Indiana	-39.90%	-27.85%	-56.64%	-06.86%
Iowa	-27.71%	-20.40%	-42.45%	-05.19%
Kansas	-28.07%	-20.74%	-42.99%	-07.10%
Maine	-69.37%	-34.90%	-80.06%	-07.82%
Maryland	-27.92%	-68.42%	-77.23%	-06.60%

(1) Permission for use granted by Karen Kerstein, ASCUS

TABLE B

NUMBER OF STUDENTS EARNING DEGREES IN EDUCATION BY FIELD IN 1978-79 FROM ONE INSTITUTION

Fields in which there are reported shortages	No.	Fields in which there are reported surpluses	No.	Other fields with more balance	No.
Business Education	12	Art	47	Miscellaneous	85
Chemistry and Science	7	Elementary Education	183		
Industrial Arts	30	P. E.	93		
Math	15	Social Studies	22		
Special Education	200				
Speech Correction	8(a)				
Total	**272**		**345**		**85**

(a.) Speech "Communication" was the category named.

Teacher Supply and Job Demand for the Early 1980's

by Leonard L. Pourchot

How many students will be in school in the 1980's?

How many teachers will be needed in

From

Place
Stamp
Here

Thresholds in Education
P.O. Box 771
DeKalb, IL 60115

increase in births continues, a slight increase in the demand for elementary teachers may be seen in the mid-1980's, but these slight increases would not be seen in high schools until after 1990.

How many teachers will be needed in the next several years?

During the 1970's, teacher production shrank to half the size of the peak year of 1971-2. Production of both elementary and secondary teachers has been cut back, but secondary teaching has experienced the more severe reduction of numbers. Apparently supply/demand fluctuates by geographical area and by subject fields. During the 1970's, severe cutbacks in teacher preparation occurred throughout the United States, but from 1978-79 to 1979-80 it appears that the situation has largely stabilized. In fact, in '79-'80 the number of total teachers prepared has

art
elementary education
health
physical education - men
physical education - women
social studies

Other subject areas show more of a balance between supply and demand.

Where in the United States are teaching jobs available?

Reports from placement directors showed widely scattered reactions, but some generalizations may be drawn. Elementary and junior high supply and demand seemed balanced except for shortages reported in scattered areas including Colorado and Wyoming.

Mathematics, natural and physical sciences, trade, industrial and vocational-technical teachers, and special education teachers

shortages were reported had comparatively few teaching candidates. If the experience of the one institution is indicative of what is happening in similar institutions, both the teacher shortages and the teacher surpluses may persist into the future.

References

1. ASCUS Research Report. **Teacher Supply/Demand in the United States 1980.** March, 1980.
2. Career Planning and Placement Center. Northern Illinois University. "Placement Data for December 1978—May/August 1979 Undergraduates Earning Degrees in Education."

TABLE A[(1)]

A COMPARISON OF PERCENTAGE CHANGE IN NUMBER OF TOTAL NEW TEACHERS PREPARED BY STATE FOR SELECT YEARS

	1971-72 vs 75-76	1975-76 vs 78-79	1971-72 vs 78-79	1978-79 vs 79-80
Alabama	-25.62%	-03.65%	-28.34%	-06.07%
California	-13.85%	-35.86%	-44.74%	-01.17%
Colorado	-46.80%	-27.33%	-61.34%	-06.42%
Florida	-13.56%	-02.36%	-14.53%	-09.82%
Georgia	-38.46%	-45.83%	-66.66%	+27.77%
Idaho	-49.53%	-12.96%	-56.07%	00.00%
Illinois	-19.97%	-35.12%	-48.08%	-07.45%
Indiana	-39.90%	-27.85%	-56.64%	-06.86%
Iowa	-27.71%	-20.40%	-42.45%	-05.19%
Kansas	-28.07%	-20.74%	-42.99%	-07.10%
Maine	-69.37%	-34.90%	-80.06%	-07.82%
Maryland	-27.92%	-68.42%	-77.23%	-06.60%
Massachusetts	-40.38%	-14.60%	-49.09%	-06.39%
Michigan	-38.56%	-19.88%	-50.78%	-17.72%
Minnesota	-43.67%	-25.33%	-57.94%	+00.98%
Missouri	-36.13%	-30.95%	-55.90%	-02.69%
Montana	-20.09%	-16.00%	-32.87%	+03.28%
Nebraska	-33.10%	-19.95%	-46.45%	-04.97%
New Jersey	-33.49%	-43.26%	-62.26%	-16.25%
New Mexico	-19.38%	-25.00%	-39.54%	-05.30%
New York	-12.87%	-25.47%	-35.06%	-14.34%
North Carolina	-08.73%	-40.96%	-46.50%	+44.30%
North Dakota	-29.96%	-26.38%	-48.44%	-04.15%
Ohio	-40.89%	-22.58%	-54.24%	-02.25%
Oregon	-31.39%	-22.99%	-47.17%	+22.82%
Pennsylvania	-43.12%	-24.49%	-57.05%	-06.95%
Rhode Island	-51.07%	-11.76%	-56.83%	+24.05%
South Dakota	-40.74%	-17.67%	-51.21%	+02.29%
Tennessee	-05.22%	-14.70%	-19.16%	+11.03%
Texas	-16.58%	-31.36%	-42.45%	00.52%
Utah	-25.38%	-05.68%	-29.62%	-04.36%
Virginia	-19.51%	-18.61%	-34.49%	-07.44%
Washington	-24.06%	-31.35%	-47.87%	-00.00%
West Virginia	+01.69%	-53.67%	-52.87%	-04.87%
Wisconsin	-16.75%	-25.51%	-37.99%	+09.53%
Wyoming	-31.06%	-21.60%	-45.95%	-03.38%

(1) Permission for use granted by Karen Kerstein, ASCUS

TABLE B

NUMBER OF STUDENTS EARNING DEGREES IN EDUCATION BY FIELD IN 1978-79 FROM ONE INSTITUTION

Fields in which there are reported shortages	No.	Fields in which there are reported surpluses	No.	Other fields with more balance	No.
Business Education	12	Art	47	Miscellaneous	85
Chemistry and Science	7	Elementary Education	183		
Industrial Arts	30	P. E.	93		
Math	15	Social Studies	22		
Special Education	200				
Speech Correction	8(a)				
Total	**272**		**345**		**85**

(a.) Speech "Communication" was the category named.

THRESHOLDS IN EDUCATION
COLLEGE OF EDUCATION
NORTHERN ILLINOIS UNIVERSITY
DEKALB, ILLINOIS 60115

RETURN POSTAGE GUARANTEED

NON PROFIT ORGANIZATION
U.S. POSTAGE PAID
PERMIT 120
DEKALB, ILLINOIS 60115