DOCUMENT RESUME

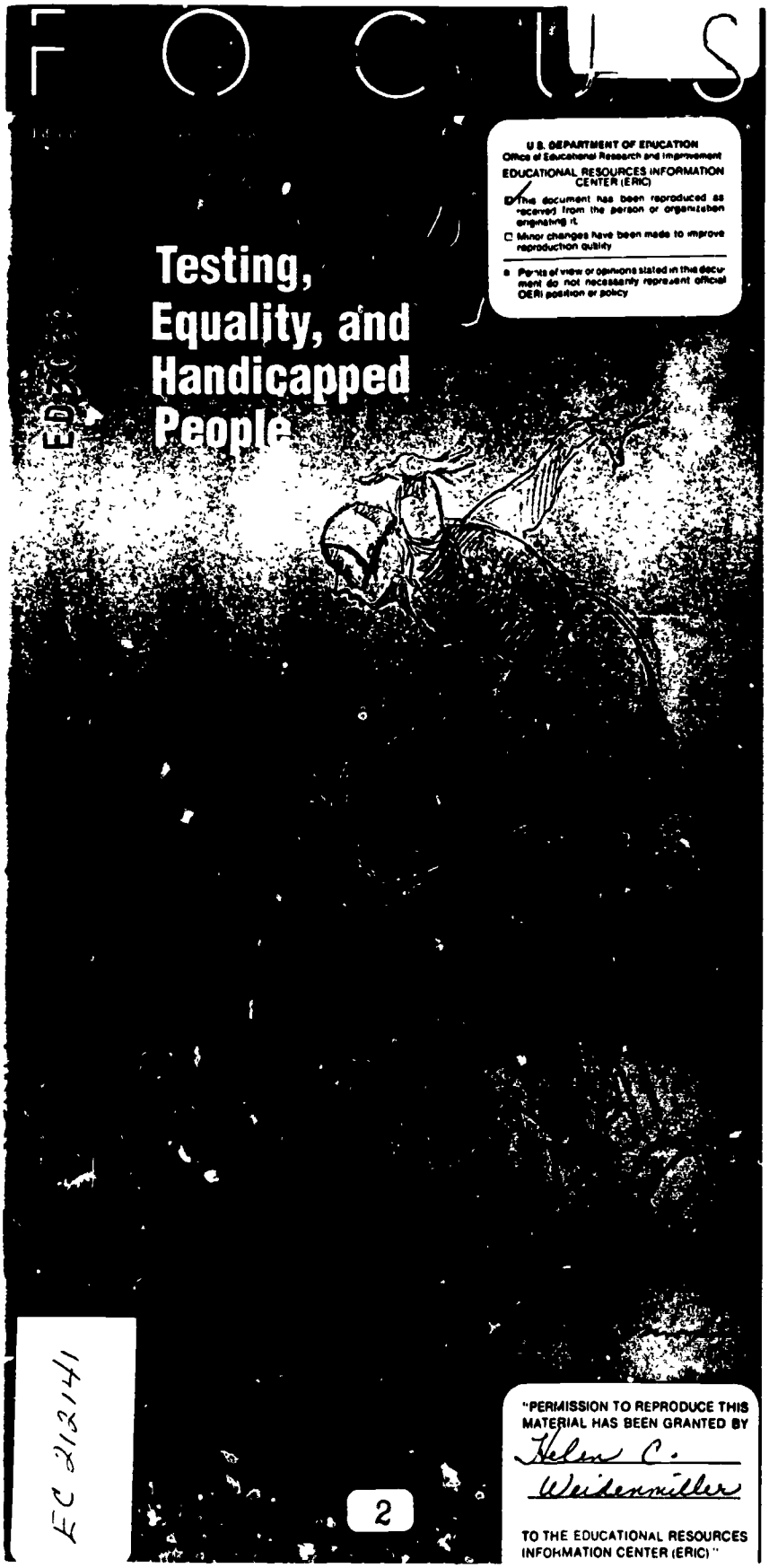
ED 303 944 EC 212 141

AUTHOR Benderson, Albert, Ed.
TITLE Testing, Equality, and Handicapped People.
INSTITUTION Educational Testing Service, Princeton, N.J.
PUB DATE 88
NOTE 23p.
AVAILABLE FROM FOCUS, Educational Testing Service, Princeton, NJ 08541-0001.
PUB TYPE Reports - Descriptive (141) -- Collected Works - Serials (022)
JOURNAL CIT Focus; v21 1988

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Admission Criteria; College Admission; *College Entrance Examinations; College Students; Comparative Analysis; *Comparative Testing; Difficulty Level; *Disabilities; Factor Structure; Higher Education; *Predictive Validity; Scaling; Selective Admission; *Testing Problems; Test Items; Test Reliability; *Test Validity

ABSTRACT

The scores of handicapped students taking tests such as the Scholastic Aptitude Test (SAT) or the Graduate Record Examinations are flagged so that admissions officers will be aware that they were achieved under special circumstances. A series of studies was initiated to determine whether special administrations of such tests are comparable to standard administrations, in which case flagging would no longer be necessary. The studies looked at comparability data for test takers with hearing impairments, visual impairments, physical handicaps, and learning disabilities. Comparability between standard and nonstandard test forms was found to be high, particularly with respect to characteristics as reliability, factor structure, and differential item difficulty. Analysis of the tests' predictive validity with regard to academic performance found that there was little over- or under-prediction for the great majority of handicapped students. The SAT did, however, substantially overpredict college performance for learning-disabled students, and this overprediction was exacerbated by time extensions during test administrations. The need for flagging test scores may be eliminated by establishing comparable timing criteria for special test administrations or by rescaling nonstandard test administrations according to how handicapped students performed in school. The comparability study also examined admissions decisions, test content, and testing accommodations. (JDD)



Testing, Equality, and Handicapped People

Standardized college admission tests were designed to provide a common yardstick for measuring the academic reasoning abilities of all students.

The Scholastic Aptitude Test, for instance, broadened college admissions by enabling students from any school anywhere in the country to demonstrate that they have academic potential equalling that of affluent students from the most elite prep schools.

Since ETS is committed to making its tests available to all students, it has traditionally made special provisions for those with handicaps. Braille and audio cassette versions are available for blind students. Special facilities are provided for those with physical disabilities. Extra time is provided for students with impaired hearing and with learning disabilities.

Unfortunately, ETS has been unable to certify that test scores earned under such special conditions are completely comparable to those taken at regular administrations. It has, therefore, traditionally flagged the scores of handicapped test takers so that admissions officers will be aware that they were achieved under special circumstances.

This practice, however, has long been the subject of considerable controversy. Advocacy groups for handicapped people have objected to flagging as a practice that identifies disabled individuals, making it easy to exclude them. The concern is that some colleges would prefer to exclude such students, thereby avoiding the expense of making special provisions for them.

Admissions officers, on the other hand, have argued that flagging is necessary if the test scores are to be evaluated accurately. They point out that disabilities can affect college performance and, therefore, must be weighed in admissions decisions.

ETS's continued use of flagging, however, has been based on its inability to guarantee the comparability of test scores.

Section 504

The passage of the federal Rehabilitation Act of 1973 intensified the controversy. Section 504 under Title V extended civil rights protection to disabled people, establishing that they are to enjoy the same protection from discrimination afforded to all other citizens. The wording of the 1977 regulations implementing the law mandated special test administrations for handicapped people while seemingly striking down the practice of flagging their scores.

One regulation, for instance, stipulates that an institution receiving federal funds must ensure that tests administered to handicapped people reflect their aptitude or achievement levels rather than their impairments.

Another regulation says that such institutions "may not make preadmission inquiry as to whether an applicant is a handicapped person."

Flagging has been viewed by some as a violation of this second regulation because a flagged score report reveals that the test taker has a disability. Nevertheless, testing organizations have been reluctant to distribute scores achieved under special circumstances without indicating that they might not be equivalent to the same scores achieved under standard testing conditions. They also feel constrained by professional standards established by the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education, recommending that users be cautious about nonstandard scores achieved when comparability is uncertain.

"Comparable scores do not necessarily imply the same average score for handicapped and non-handicapped groups..."

To resolve this dilemma, the National Academy of Sciences was asked to impanel a committee to reconcile the testing requirements of the new law with sound psychometric practice. In its 1982 report, the panel agreed that "current psychometric theory and practice do not allow full compliance with the regulations as currently drafted."

The panel recommended, therefore, that a four-year study be conducted to determine whether tests modified for handicapped test takers are comparable to standard versions and whether they provide accurate estimates of the academic ability of students with disabilities. If the predictive validity (or accuracy) of both versions were found to be comparable, the panel suggested, it would no longer be necessary to flag the scores of handicapped test takers.

In response, Educational Testing Service, together with the College Board and the Graduate Record Examinations Board, initiated in 1983 a series of pioneer studies to determine whether special administrations of the Scholastic Aptitude Test and the Graduate Record Examinations for handicapped students are comparable to standard administrations.

The project culminated in March 1988 with the publication by Allyn and Bacon, Inc., of *Testing Handicapped People* by Warren W. Willingham, Marjorie Ragosta, Randy Elliot Bennett, Henry Braun, Donald A. Rock, and Donald E. Powers. The book provides, for the first time, answers to some of the most vexing questions surrounding the comparability of scores and offers a series of recommendations for the future.

The studies looked at comparability data for all four categories of disability:

- hearing impairment, which ranges from hard of hearing to deafness;
- visual impairment, which may range from a serious visual deficit to blindness;
- physical handicap, which includes a variety of neurological and orthopedic disabilities;
- learning disability, defined as a specific perceptual, neurological, or cognitive deficit identified mainly on the basis of school achievement.

The focus was on three questions for each of these groups.

1. When admission tests are modified for handicapped people, to what extent are the nonstandard tests and the resulting scores comparable to those of the regular national program?
2. Might the comparability of such tests be improved? If so, how?
3. What implications might be drawn for possible resolution of the flagging problem?

Researchers were concerned with both score comparability and task comparability. If the scores of handicapped test takers are comparable, they will reflect only aptitude or ability, rather than extraneous limitations or impairments. The test must measure the same factors as the standard examination, and it must predict college performance as accurately as the standard test.

The fact that all scores studied were flagged complicated the research task because it was entirely possible that the flags themselves affected admissions decisions. If handicapped students were admitted on a fundamentally different basis than nonhandicapped students because of the flags, as critics allege, the comparability of scores, particularly in terms of predictive validity, would be more difficult to determine.

With respect to task comparability, the cognitive demands of the test must be shown to be equivalent for handicapped and nonhandicapped test takers. The content must be comparable, no matter how it is presented; the accommodations must be appropriate, and the timing must be equivalent, even if handicapped students are allowed additional time to complete test questions.

Willingham writes, "The matter often comes down, in the last analysis, to a judgment about what is reasonable and fair in testing people with a particular disabling condition."

Researchers established an exhaustive series of criteria to be used in determining whether special administrations of the Scholastic Aptitude Test (SAT) and Graduate Record Examinations (GRE) were comparable to standard administrations. According to Willingham, these included the performance of handicapped students on different types of test materials or formats, the performance in college of nonstandard test takers, evidence of the speed with which handicapped students complete tests, and the comparability of handicapped student accommodations in admissions testing to those used in college testing.

Researchers tracked the performance of students with each of the four handicapping conditions on most versions of the tests to determine how differing formats affected performance. The frequency with which different groups completed tests with different time limits was tabulated. Researchers were also concerned with the reliability or precision of test scores and with whether specific test items measured the same factors for handicapped and nonhandicapped students.

Test results were tallied for most configurations. The scores for visually impaired students, for instance, were tracked for regular-type, large-type, and braille editions of the SAT. Results were also compared to those achieved by regular students using standard test forms.

Sophisticated statistical measures were applied to test questions to determine whether they were measuring the same factors for all test populations.

Finally, SAT and GRE scores were correlated with first-year grades to determine whether special versions of tests administered to handicapped students predicted their performance in college with accuracy comparable to that yielded by standard versions.

E

Willingham cautions that differing versions of a test do not have to be identical for them to be comparable. "Comparable scores," he writes, "do not necessarily imply the same average score for handicapped and nonhandicapped groups because there is no way to know whether the groups are either representative of students generally or comparable in their learning experience.

"The important objective," he adds, "is to make the task as comparable as possible by removing irrelevant sources of difficulty."

Three processes, he says, are involved in answering test questions — sensory-motor, encoding, and higher-level cognitive proc-

esses College admission tests are designed to measure cognitive abilities, and interference arising from defects in other processes must be screened out The defective sensory-motor processes of those with physical handicaps and the limited encoding processes of blind and learning disabled students must in no way affect test outcomes if results are to be considered comparable

The researchers measured comparability in eight dimensions reliability, factor structure, differential item difficulty, prediction of academic performance, admissions decisions, test content, testing accommodations, and test timing

Overall, comparability between standard and nonstandard test forms was found to be high, particularly with respect to such internal characteristics as reliability, factor structure, and differential item functioning. Test results were not affected by the extraneous physical limitations of handicapped test takers.

Across the board, for instance, the tests were found to be highly comparable with respect to reliability — their measurements are equally precise for handicapped and nonhandicapped test takers.

Willingham points out, however, that it must also be demonstrated that the tests measure the same thing. Factor analysis is the statistical method used to make this determination.

Except for the fact that verbal and quantitative abilities were found to be less closely related for handicapped test takers, the factor analysis revealed test forms to be highly comparable. Willingham writes, "The similarity in the tests' factor structure for handicapped and nonhandicapped examinees supports the assumption that the nonstandard test scores represent comparable cognitive abilities and that they have not been distorted by the student's disability."

Although nonstandard and standard tests were shown to measure comparable cognitive abilities, it remained possible that nonstandard versions might contain questions that were inappropriate because they were particularly difficult only for disabled test takers A differential item functioning analysis was conducted to determine whether such bias existed, and except for a few questions on the braille version of the mathematical portion of the SAT, little evidence of such questions surfaced

With respect to test content, it would seem self-evident that tests delivered in standard and nonstandard administrations must be comparable since the content is identical The issue, however, is not necessarily so easy to resolve, because identical questions might not be comparable if they are made more difficult by the disabilities of some test takers

Although such problems were found to be rare, the report questions the comparability of SAT and GRE verbal questions for hearing-impaired students Those who have been deaf from birth have particular difficulties communicating in or understanding written English, which is a fundamentally different language from the sign language they normally use to communicate. For instance, the various forms of sign language typically lack articles and prepositions, and their grammatical structures differ radically from English. Students who have never heard English have an extremely difficult time comprehending its structure or meaning.

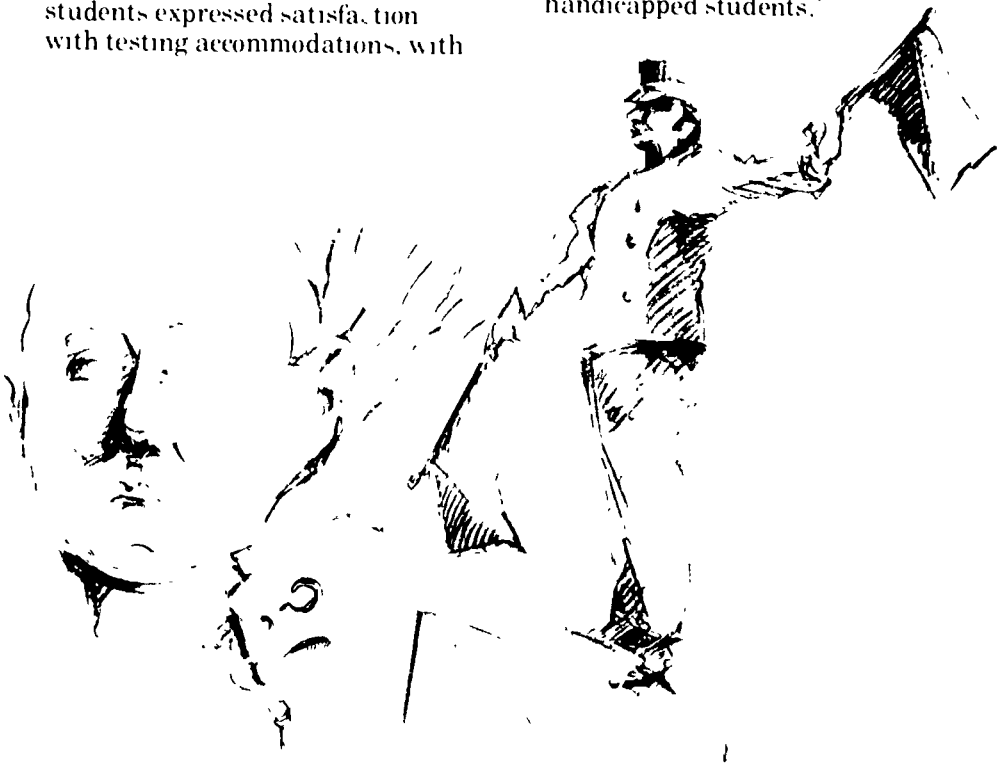
These difficulties were reflected in the average SAT verbal and mathematical ability scores for hearing-impaired students, which were considerably lower than those achieved by other handicapped groups. The report suggests that for some deaf students, these low scores may reflect the noncomparability of test questions rendered unnecessarily difficult by the students' lack of English communication skills. Willingham points out that manually fluent students tended to receive the lowest scores. He suggests, therefore, that a sign-language version of the test might provide a more valid assessment of their skills and recommends that the feasibility of such a test should be examined.

An investigation of admissions decisions revealed that, contrary to the assumptions of those in some handicapped advocacy groups, the selection process for handicapped applicants was generally comparable to that for nonhandicapped students. Despite flagged scores, admissions of handicapped and nonhandicapped students alike increased in direct proportion to increases in their high school grades and test scores. The effect of flagging, therefore, seemed minimal.

A high percentage of disabled students expressed satisfaction with testing accommodations, with 94 percent of SAT test takers and 86 percent of GRE test takers approving of testing conditions.

Establishing the predictive validity of the tests with regard to the academic performance of disabled students was absolutely essential, Willingham explains. "The NAS panel viewed the accuracy of grade predictions as a crucial aspect of comparability, and with good reason. The validity issue is whether one can safely make the same inference as to future academic performance when looking at test scores from nonstandard and regular administrations. Do the nonstandard scores predict performance accurately? Are they useful to the college and fair to the students?"

Test scores for handicapped students taking the SAT and GRE in a variety of configurations were compared to first-year grades in college and graduate school. As with nonhandicapped students, the accuracy of predictions was enhanced by combining test scores with high school grades. The report concludes that "when academic performance was predicted on the basis of test scores and prior grades, there was little over- or underprediction for the great majority of handicapped students."



Willingham writes, "This is an important finding because it indicates that if admissions officers follow the standard advice and usual practice of using grades as well as test scores in estimating future performance, these estimates will not, on average, be either too high or too low."

When looking at subgroups of test takers, rather than aggregate groups, however, researchers found that the academic performance of handicapped students was less predictable than that of their nonhandicapped classmates. This finding held up whether test scores, grades, or both were used as predictive instruments, and it was applicable at both undergraduate and graduate levels.

"If you break down the group into high and low scorers on any predictive measure," says Willingham, "the handicapped students were less predictable. Those who score quite high on the test do worse in school than you would expect, and those who score low do better."

Willingham attributes some of this lower predictability to variations in the quality of educational programs for the disabled and to outside factors, such as financial problems or lack of support programs, that have a particular impact on those with handicaps.

Results also varied for different handicaps. For instance, test scores substantially underpredicted college grades for hearing-impaired students enrolled in college programs that provided special services for them. That is, grades in these programs were higher than test scores predicted. (When the college performance of hearing-impaired students in regular college programs was predicted on the basis of tests and grades, accuracy was high.)

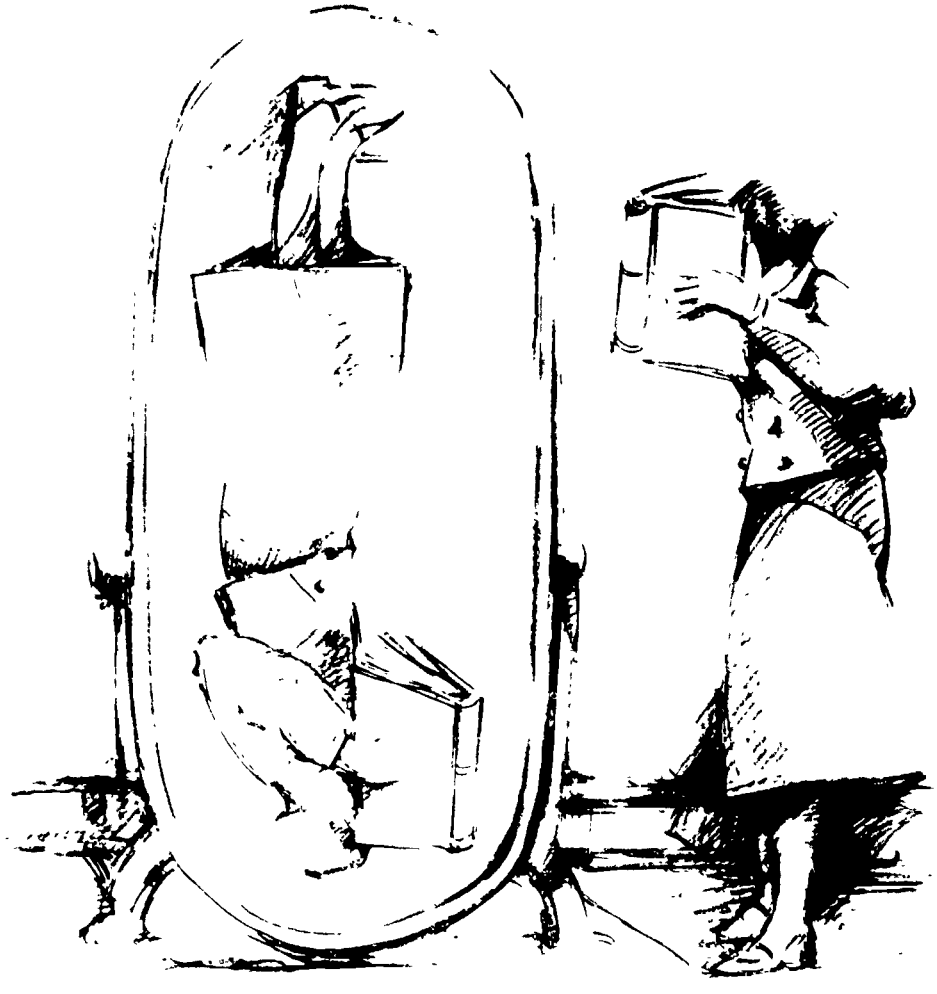
On the other hand, the SAT overpredicted college performance for both the physically handicapped and the learning disabled. For the physically handicapped, the overprediction was not very large, but for the learning disabled, the degree of overprediction was substantial. Although these students were not significantly overpredicted when grades were added to test scores, because high school grades were significantly lower, the result was still troubling.

Willingham points out an inherent problem in establishing the predictive validity of scores for learning disabled students. People are identified as learning disabled precisely because their achievement does not measure up to their test scores. One of the primary criteria for distinguishing learning-disabled from slow learners is precisely the fact that while they do well on tests of ability and have average to above-average I.Q.s, their academic performance does not measure up to these test results.

In contrast to disabilities arising from physical deficiencies and readily apparent to observers, learning disabilities are primarily academic disabilities. They are defined by poor academic performance and cover a wide range of conditions including dyslexia, perceptual handicaps, and minimal brain dysfunction. Such conditions are not readily apparent to observers, and diagnosis can be highly subjective.

The Education for All Handicapped Children Act of 1975 defined learning disability as "a disorder in one or more of the basic psychological processes involved in understanding or using language, spoken or written, which may manifest itself in an imperfect ability to listen, think, speak, read, write, spell, or do mathematical calculations." The definition specifically excludes visual, hearing, or motor handicaps, mental retardation, and the effects of environmental, cultural, or economic disadvantage.

One of the most common learning disabilities is dyslexia, a condition characterized by impaired ability to read. Dyslexics may transpose letters in words, mistake one word for another, skip words or lines entirely, or have difficulty sounding out words.



Learning disabilities may also include other language processing problems such as short-term memory deficits that render readers incapable of remembering what they have just read and organizational deficiencies that make it impossible for readers to distinguish main ideas from supporting evidence. Although many of these problems seem to be physiologically based, the exact mechanisms at work are unclear pending further research.

Willingham points out that federal regulations specify that learning disabled students be identified on the basis of poor school performance in relation to ability. "It would seem to make little sense," he writes, "to evaluate score comparability for this group on the basis of over- or underprediction (from test scores alone) when a discrepancy in the test-school achievement relationship is precisely the basis upon which the group is identified!"

Two factors seem to underlie the overpredictive scores of learning-disabled students. The first is the imprecise definition of learning disabilities, and the second is the time allowed to complete the test.

The numbers of people in this category have grown tremendously in recent years, and today more than 1.8 million pupils are identified as learning disabled. There is strong evidence that some of this growth reflects the tendency of some schools to place many students without physiologically-based learning problems into this category inappropriately.

"There's a lot of social and educational cleavage on how to view all this," says Willingham. "When we see enormous increases in the numbers of people identified as learning disabled, many suspect some kind of educational game-playing. Some suspect that schools might funnel more people into this category to attract more program funds, rather than maintaining an accurate scientific or educational categorization. Whereas a lot of people used to be labeled mentally retarded, now fewer are placed in that category and more are labeled learning disabled. Since learning disabilities are not clearly labeled sensory deficits, this makes people skeptical and arouses controversy."

In 1982, Lorrie A. Shepard and Mary Lee Smith, both from the University of Colorado Department of Education, conducted a study of learning-disability placements in the State of Colorado. The results were summarized in their Spring 1983 *Learning Disability Quarterly* article "An Evaluation of the Identification of Learning Disabled Students in Colorado." They wrote, "Approximately 60 percent of the pupils currently identified as LD do not match the legal definitions or the definitions presented in the professional literature."

Shepard and Smith found that many students had been inappropriately placed in the learning-disabled category. They found that 26.8 percent had been placed in learning-disabled classes without any I.Q. test data, 28.5 percent had I.Q.s below 90, and 8.3 percent had I.Q.s below 80.

Among the student population identified by the schools as learning disabled, only 43 percent demonstrated actual learning disabilities. The remainder included slow learners, emotionally disturbed students, and non-native English speakers.

If Shepard's findings hold true for the nation as a whole, they imply that scores on the special administrations of the SAT for the learning disabled may be artificially inflated when additional time is provided to students who are inappropriately placed in that category.

In fact, the researchers found that the tendency of standardized tests to overpredict academic performance for the learning disabled was exacerbated by time extensions. Willingham reports that providing extended time to learning-disabled test takers "may raise scores beyond the level appropriate to compensate for the disability."

Learning-disabled students are currently allowed up to 12 hours to complete the SAT, a virtually unlimited block of time. Learning-disabled students who took the most time to complete the test earned the scores that most seriously overestimated college performance. These results have raised serious questions about whether or not learning-disabled students should be granted additional time to complete exams and, if so, how much extra time should be granted.

Generally, all handicapped students, regardless of disability, achieved higher scores when they took additional time on the exam. Increases ranged from 30 points for the hearing impaired to 38 for the learning disabled. Taking additional time also increased the chances of reaching late items and answering them correctly, even though the final items tend to be the most difficult.

However, the first-year college performance of learning-disabled students, unlike other groups, was significantly overpredicted when they took additional time, and the degree of overprediction increased the more time they took. Willingham writes, "This appears to be direct evidence that the SAT scores of LD students who took longer amounts of time on the test were somewhat inflated."

There were also lesser indications that timing may inflate scores for physically handicapped and hearing impaired test takers, but in both cases the effects were minimal or vitiated by other factors. Nonetheless, these findings led Willingham to conclude that timing represented the only aspect of nonstandard test administrations that was not comparable to standard admini-

strations. He says that scores are raised at least somewhat beyond the level that would be achieved with comparable time, although the problem is acute only with respect to test takers who are learning disabled.

These results have led some to question the philosophical basis for allowing learning-disabled students to have large amounts of additional time to complete a test such as the SAT or the GRE.

Marjorie Ragosta says that because learning-disabled students achieve relatively low scores on college admission tests, research must be done to determine whether these scores are due to inaccurate measurement or are an accurate reflection of the students' achievement levels. "If the differential performance is not caused by measurement inaccuracies," she says, "does it make sense to turn around and say that you have to adjust the methods of testing or give a different test because it reflects differential performance? Is it realistic to try to test people with a learning disability as if they don't have it?"

Ragosta speculates that overprediction arises on the SAT because some learning-disabled students — particularly those who take seven or eight hours extra — are receiving relatively more time for the test than they can continually give to their college assignments. "What we are doing, in effect, for one little part of this individual's life, is to allow unlimited time that is not feasible everywhere," she says.

Willingham says, "To label students learning disabled and say they should have more time on the test because they don't do well at tests seems to stand the argument on its head."

Randy Bennett suggests, however, that overprediction may also be caused by the fact that it takes learning-disabled students longer to become oriented to college. "High school special education programs are highly structured, while college is not. Learning-disabled students, he argues, might have more difficulty adjusting in the first year and then do better in subsequent years.

He also suggests that the overprediction reported in the study may not hold for those attending colleges with special programs for

learning-disabled students. As indicated earlier, test scores, in fact, underpredict the college grades of learning-disabled students in special programs. Such programs provide help with study skills and often permit students to take a lighter course load each semester. This, in effect, allows them to devote additional time to each subject. He points out that in the last few years the number of these programs has increased significantly, so that even the most prestigious schools, such as Dartmouth and Brown, now have special programs for learning-disabled students.

"If we don't give extra time to students who will be in special programs in college," says Bennett, "the test will be just as invalid as if we provide extra time to students who receive no extra help in college."

Sally Shaywitz of the Yale Medical School says that the need for extra time is fundamental to the definition of learning disabilities. "Learning-disabled students don't need extra programs, but they need more time to process information and get it on paper," she says. "The whole discrepancy is between their intelligence and what they can do in a given amount of time. For learning-disabled students, extra time is absolutely crucial. To deny it is the kiss of death.

"If you don't know what kind of accommodation the kids had at college, you can't determine the predictive validity of the test," she adds.

Willingham, however, disputes the notion that all learning-disabled students need extended time on admission tests. He suggests that ETS data indicate that only seriously disabled students need extra time to obtain a score that has predictive validity. For most learning-disabled students, he says, extra time merely leads to an inflated prediction of college performance.

"These results," he writes, "suggest that testing programs need to reevaluate their policies regarding extended time for LD students, especially as to how much time should be allowed and whether it is possible to improve present practices concerning eligibility for the nonstandard examination."

"For learning-disabled students, extra time is absolutely critical. To deny it is the kiss of death."

Ragosta suggests that standards be tightened so that only students who can demonstrate a history of accommodations for a learning disability throughout their education be permitted to have extra time on the SAT or GRE. Currently, students can qualify for a special administration by presenting two pieces of documentation from experts in learning disabilities or evidence of an Individualized Education Program. All special education students in elementary and secondary schools are supposed to have such a program setting individual educational goals. It is developed by a committee of school officials (including a teacher) in cooperation with the child's parents and should reflect a realistic assessment of what the child can learn and what kind of special help will be needed.

Ragosta expresses skepticism about some expert documentation and suggests that Individualized Education Programs should be the primary qualification for extra help for people in public school systems. Assistance on the test would reflect the assistance received at school. For instance, visually handicapped students using large-print texts in school would receive large-print tests. Similarly, learning disabled students receiving extra time to do school assignments would be allowed extra time to complete the SAT or GRE.

"I don't think it's fair for the testing company to assume the entire burden of deciding how students will be tested," she says. "We have only one encounter with the individual and don't really know what his or her disability is. If the Individual Education Program says that a student should take the test only with unlimited time, I think that's a good indication that the proper accommodations should be."

In lieu of [illegible] gent qualifying criteria [illegible] ial administrations, the report recommends that time limits be established for all handicapped groups comparable to the time limits imposed on nonhandicapped test takers. Since the SAT and GRE have traditionally set

time limits deemed adequate for 80 percent of test takers to answer all questions, it has been recommended that the same standard be established for handicapped test takers.

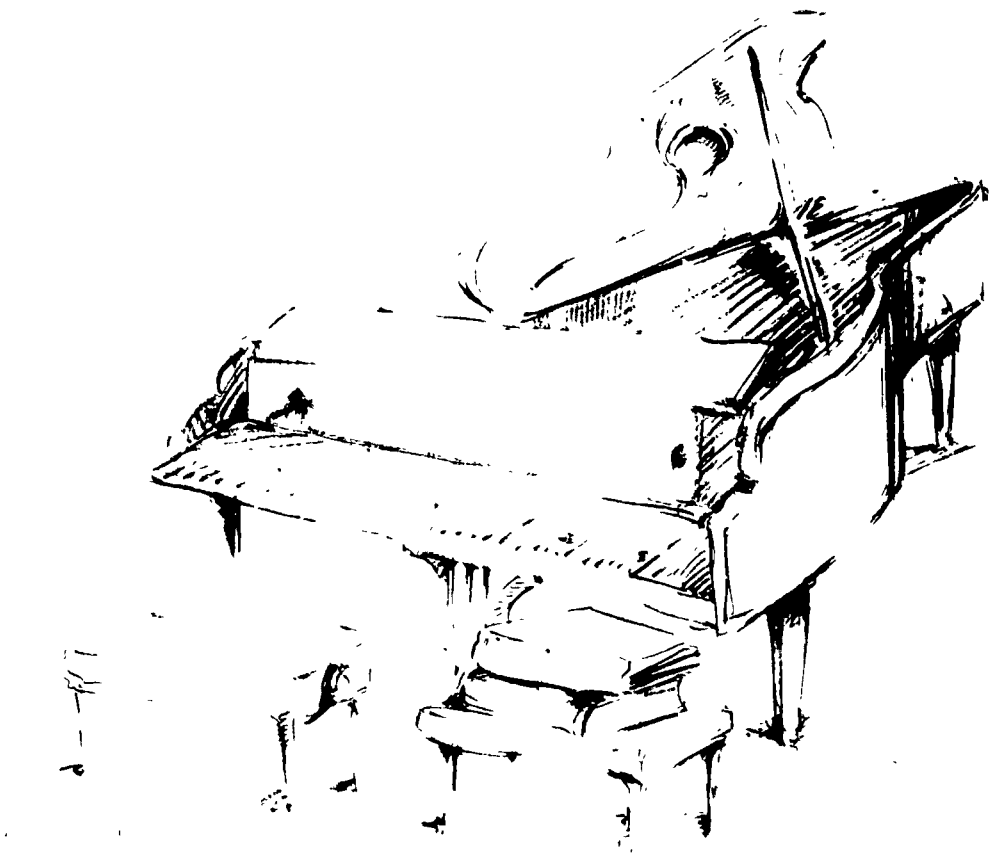
In order to carry out this recommendation, Ragosta has embarked upon a study to determine the time it takes 80 percent of those in each handicapped category to complete the SAT and the GRE. Presumably, this will make the timing on special administrations more comparable to that on regular administrations and will help to alleviate the overprediction problem.

In order to determine these time limits, Ragosta will review timing records compiled by the various handicapped groups during past years.

"When we provide 12 hours, which is a virtually unlimited amount of time, we are saying that every handicapped student should have a chance to finish the exam," says Ragosta. "There should be at least some equality to be fair to the population at large. If we can determine the time that allows 80 percent of those with a disability to finish, that should provide a cut time comparable to that given the general population.

"Of course, we want to leave a loophole for those for whom the severity of disability precludes meeting this standard."

In order to strengthen knowledge of predictive validity, Ragosta and ETS research scientist Henry Braun have also initiated a study to compare handicapped students' test scores to four-year, rather than first-year, college performance. It is hoped that the study will yield a more solid estimate of the predictive validity of special administrations. The researchers will also attempt to discern whether handicapped students generally take longer to complete college than do nonhandicapped students, an issue they consider relevant to the granting of additional time on standardized tests.



Bennett expresses the hope that the study of four-year progress in college will also reflect the impact of special programs for the learning disabled recently instituted at many colleges.

In addition, Bennett is also conducting a study of item bias for students with visual handicaps on the SAT mathematics section. He is attempting to discover which types of items don't work for blind students. Preliminary results indicate that the abilities of blind students cannot be tested accurately by items containing drawings and small diagrams and by those that ask test takers to estimate solutions based on visual material.

R

It's possible that establishing comparable timing criteria for special test administrations will help solve the ongoing dispute over the practice of flagging test scores. Handicapped advocacy groups have long pressed for an end to flagging because they fear that the practice provides an easy method for spotting and rejecting the applications of handicapped students. Until now, ETS has flagged handicapped students' test scores because it could not guarantee their comparability and therefore was bound by established professional standards.

Now that the issue of comparability has been thoroughly examined, and only relatively limited areas of noncomparability have been found to exist, an end to the practice of flagging scores seems within sight. Establishing comparable timing limits for special test administrations would go a long way towards solving the problem. Another possibility, also investigated, would be to rescale the nonstandard administrations according to how handicapped students performed in college and graduate school. Rescaling might also eliminate the need for flagging, perhaps without limiting test-taking time.

Rescaling was first suggested by the National Academy of Sciences panel. It proposed that scores of handicapped test takers could be made to predict college performance with the same degree of accuracy as those of nonhandicapped students by adjusting the scores according to some sort of statistical formula. For instance, scores might be adjusted according to how handicapped students performed in school so that an 800, for instance, would represent the highest level of work students with a particular handicap accomplish in higher education.

Donald E. Powers and Willingham conducted an extensive study of the rescaling proposal. Powers says, "The proposal was that you could make the scores of handicapped and nonhandicapped test takers comparable by looking at how both types of students perform during the first year of school, taking the first-year grade-point average as a common link, and then adjusting test scores to obtain a comparable prediction of first-year performance. It would entail adding a constant to modify the scores of handicapped students. It seemed like a proposal worthy of consideration."

Ultimately, however, the researchers rejected that proposal. They found that it was not possible, given the limited number of people with various degrees of handicaps, to collect a large enough pool of data upon which to base scaling decisions. The problem was particularly acute for the GRE General Test, which is taken by far fewer students than the SAT.

Moreover, they concluded that grade-point averages would not provide a sufficiently reliable and comparable criterion for rescaling a test. The standards for grades vary widely at different colleges and this variation may be exaggerated by differences in the evaluation of handicapped and nonhandicapped students. There would be no assurance that grade-point averages as a criterion would be comparable for handicapped and nonhandicapped students, and finding national points of reference would be virtually impossible.

Willingham and Powers also found that adding a constant to the scores of handicapped students would not result in adequate scaling due to complicated variations in the predictive validity of scores for handicapped students at different ability levels. Several adjusted scores might be necessary for handicapped students, and the researchers point out that the process would be so apparent that it would be tantamount to flagging.

Other, nontechnical problems also surfaced when the rescaling proposal was examined. It might be argued that if scores were to be rescaled for one subgroup, why not for all? Racial and ethnic minorities and women, for instance, might also demand that their scores be rescaled so that differences in predictive validity, if any, between these groups and White males will be eliminated

Powers also points out that rescaling might actually harm, rather than help, some groups, such as the learning disabled. He says, "It looked like what we would have to do to adjust scores would hurt learning-disabled students because their test scores tend to overpredict grades. Downward adjustments would be hard to defend, especially in light of sparse data and the resultant shaky statistics."

He concludes, "Although, in principle, rescaling seemed not to be unreasonable, the more we looked at the data, the less technically feasible it seemed. It had a definite potential for adding inaccuracies to the system and would potentially do more harm than good. We concluded that rescaling was not a feasible way to get out of the flagging dilemma."

The unacceptability of rescaling as an alternative to flagging leaves

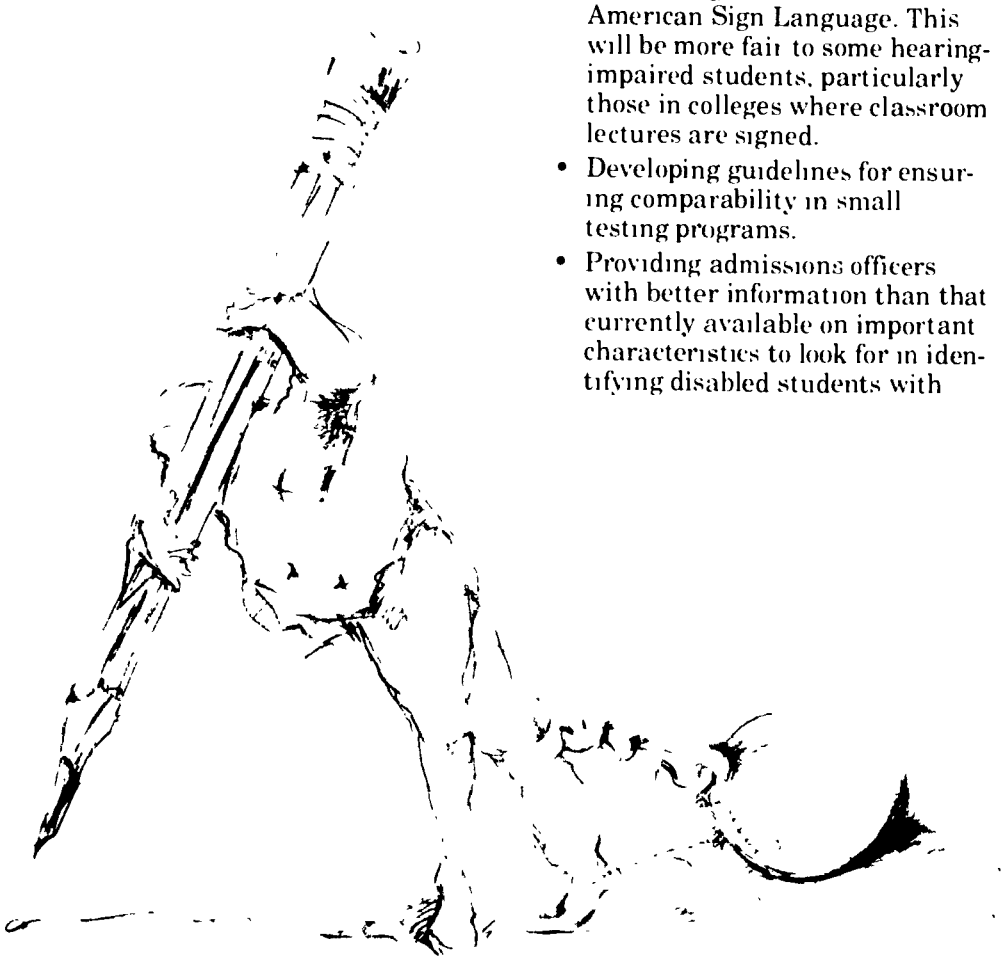
establishment of comparable time limits for handicapped students as the best hope for eliminating the practice while assuring admissions officers that the resultant scores will be comparable.

"When timing is comparable for disabled test takers," says Willingham, "they will be as likely, on average, to finish the test as nondisabled candidates. We now know that standard and nonstandard test administrations are comparable except for timing. In theory this is correctable. We should gather the data to make timing comparable and then take the flags off the scores."

L

Testing Handicapped People concludes with a number of additional recommendations for improving testing services for handicapped people. Key recommendations include:

- Routinely checking all forms of the SAT and GRE for items that may be differentially difficult for handicapped test takers. These may include items that require visual or hearing experience to understand.
- Insuring that test-familiarization and practice materials are available in braille, large type, and audio cassette formats.
- Identifying types of mathematics items that cause particular difficulty for blind students. Experience indicates that items involving three-dimensional figures may be inordinately difficult. As described earlier, a study of this problem is currently under way.
- Examining the possibility of translating admission tests into American Sign Language. This will be more fair to some hearing-impaired students, particularly those in colleges where classroom lectures are signed.
- Developing guidelines for ensuring comparability in small testing programs.
- Providing admissions officers with better information than that currently available on important characteristics to look for in identifying disabled students with



academic promise. Since the college performance of students with disabilities is less predictable from test scores and previous grades, special materials on disabilities and score interpretation might be provided to admissions officers. They should be cautioned to give less weight to traditional predictors and to take special care to review the background and personal characteristics of handicapped applicants.

- Developing better means of assessing the educational needs of handicapped students and monitoring their progress. Computer-based programs being developed by the College Board and ETS to diagnose learning problems in mathematics, writing, reading, and study skills might be adapted for handicapped students.

Generally, the results of the research reported in *Testing Handicapped People* have been encouraging. Except for the timing problem, the comparability of results on nonstandard test administrations is strong. Efforts are currently under way to solve the timing problem and allow flagging to be eliminated.

The fact remains, however, that there may never be a foolproof way to completely disentangle the effects of some disabilities from the assessment of verbal and quantitative reasoning skills by admission tests. Some disabled students, therefore, will always score lower on these measures as a result of their handicaps. The crucial question is how accurately these scores predict college performance and how comparable they are to scores attained by nonhandicapped students. The research described in *Testing Handicapped People* suggests that the comparability of standardized test results, for most handicapped test takers, remains remarkably high and will be improved in the future.

A Landmark College

ETS researchers, in *Testing Handicapped People* assert that special administrations of the SAT for learning-disabled students overpredict their performance in college in part because they are not allowed extra time to complete college assignments comparable to the extra time they receive on the test. Thus, their college grades suffer in comparison to their test scores.

There is, however, one college where the entire curriculum is designed for students with the most common learning disability — dyslexia. Just as Gallaudet University accommodates the special needs of deaf students, Landmark College in Putney, Vermont, has modified its curriculum so that it can be mastered by dyslexic students.

The school, which opened in September 1985, offers a precollege program to prepare students for college-level work and a two-year curriculum leading to an associate's degree in general studies. Its modern campus was designed by Edward Durrell Stone to house Windham College, which closed in 1978.

This year, 115 students were enrolled in the precollege program and 30 in the college division. The college program is certified by the State of Vermont Board of Higher Education and is a candidate for accreditation by the New England Association of Schools and Colleges.

Classes in the liberal arts program meet for five hours per week, rather than the traditional three. The additional class hours allow more time for course material to be presented. The instructor can also use the additional time to help students develop the study skills needed to assimilate the course material.

In addition to the regular load of three or four courses per semester, all students take a one-hour, one-to-one tutorial every other day with a faculty member who helps them with classwork, language skills, and organizational skills.

Unlike many of the special programs for dyslexics that have sprung up at colleges and universities throughout the country, the Landmark program forces students to master, rather than bypass, language skills essential to college work. All students are screened with a battery of tests before being admitted to the college, and those with extreme deficits in their language skills must first enter the precollege program to raise their skills to the twelfth-grade level. The screening process also ensures that they meet the definition of learning-disabled students, with average to above-average academic abilities, performance difficulties, and high motivation to attempt the program.

Once in the college program, however, students are not allowed to use any of the compensatory measures, such as taped books, note takers, oral examinations, and scribes, permitted at most colleges with special programs for dyslexic students. Instead, they are expected to develop the study skills necessary to engage in college-level work.

Amy Russian, assistant to the president, says, "Our program is highly competitive with a standard curriculum that is not watered down. In order to maintain the caliber of the college and keep standards high, most students must first take the precollege program. Seventy percent of the precollege students have been accepted at other colleges, but have not yet been accepted by our college program.

"Our college students have learned reading techniques to minimize the transposition of letters," she says. "However, they still may read slowly, transpose words, have spelling problems, and find writing difficult."

Russian explains that the basic work associated with teaching dyslexic students, such as overcoming letter reversals, is covered in the precollege program. In the college program, students learn more sophisticated skills, such as writing summaries to help process textbook material, advanced note-taking techniques for organizing and assimilating classroom lecture material, creating personal study guides, and organizational tools for expository writing.

Both the precollege and college programs are highly individualized, with an average class size of only six students. Individual meetings help prevent students from falling behind. All faculty have extensive training in the teaching of dyslexic students, in addition to their subject-matter training.

"It is a commonly held belief that dyslexia is neurological," says Russian. "We can't cure it, but we can teach our students to function successfully in a rigorous academic environment."