The Sixth Annual Florida Invitational Conference on Testing was held at Cocoa Beach, January 24-26, 1963. Three invited papers bearing on the Conference theme, "What's Right with Testing," were presented. Mr. Henry Chauncey, President, Educational Testing Service, Princeton, New Jersey, addressed the group and the title of his remarks was "What's Right with Testing." Mr. B. Frank Brown, Principal, Melbourne High School, Melbourne, Florida, and Mr. John R. Hills, Director, Testing and Guidance, Regents of the University System of Georgia, presented reviews of Testing, Testing, Testing, a monograph which was distributed in 1962 and was prepared by the Joint Committee on Testing established by the American Association of School Administrators, the Council of Chief State School Officers, and the National Association of Secondary-School Principals. These three papers, in their entirety, appear below.

WHAT'S RIGHT WITH TESTING

Henry Chauncey
President, Educational Testing Service
Princeton, New Jersey

I am delighted to have the opportunity of meeting with you today and saying a few words about "What's Right with Testing." We in the field of testing are working so hard on the many things that need to be done that we spend relatively little time telling the public about the important contributions tests are making to education and to our society. We are all convinced of the values of testing, but we tend to forget that others are not necessarily so informed or convinced.

Recently there has been a rash of criticisms of testing and the critics have taken to writing books instead of articles. Furthermore, one of these critics has retained a publicity agent to get him the maximum coverage in newspapers and opportunities to speak on television and radio. As a result there have been a good many occasions recently where we all have been subjected to the slings and arrows of outrageous attacks.

Now there are two things that I would say. First, the critics of testing are not representative of the general public. They tend to be individuals who have a certain combination of verbal fluency and personality disposition that results in their rushing into print. They,

therefore, are more audible and more visible than other persons who recognize the values of testing and take tests as a natural and important part of our society. Secondly, the reviews of the critics' books, of which there have been a good many, have in almost every case been unfavorable to the books and have recognized the need for and the value of tests.

The general acceptance of tests is documented by the results of two recent opinion polls. Last November the Gallup Poll posed the following question to a representative cross section of the American public:

> Would you favor or oppose a nationwide test to be given to all high school students when they enter and when they graduate to enable parents to know more about their children and their educational achievements?

Whatever we may think about the problems involved in such a proposal, it is interesting that 77 per cent answered yes, 13 per cent answered no, and 10 per cent had no opinion.

About a year ago, just at the time of the publication of Testing, Testing, Testing, the NEA Research Division reported in the NEA Journal the response to a question asked in the Teacher Opinion Poll. The question asked was:

> Are nationwide testing programs, such as merit scholarship examinations and college entrance examinations, exerting an influence on the instructional program of your school?

Of the elementary school teachers, 39 per cent said yes, 36 per cent said no, and 25 per cent were undecided. Of the secondary school teachers, 56 per cent said yes, 28 per cent said no, and 16 per cent were undecided.

The teachers answering "yes" to this question were asked whether they considered the influence of national testing programs desirable or undesirable. An overwhelming majority of these teachers (88 per cent) considered the influence desirable.

Even though the large majority of the general public and of teachers favor the use of tests and recognize the important contributions that

tests have made, there are a number of reasons why, in some quarters, tests are unpopular and the object of hostility. John Gardner in his book, Excellence, mentions several.

.... "Many people," he says, "have an aversion to being the subject of mental diagnosis."

.... "Many fear that tests will be inaccurate" and that they will not provide a fair appraisal.

.... "Apprehension is fostered by the fact that ... the processes of mental measurement" are hard to understand, and "no one wishes to be judged by a process that he cannot comprehend."

.... There is also the "fear of the potentialities for social manipulation and control inherent in any large-scale processing of individuals."

"Tests," Gardner says, "are designed to do an unpopular job." He goes on to point out that even if tests were improved and did become "less vulnerable to criticism," the hostility to them would actually increase!

So it is not surprising that there are some very vocal critics of testing. And, because an attack on something is always news, it is not surprising that they get more than their fair share of space in the public press. I think we have to expect and accept, with as much patience as possible, periodic criticisms of tests.

This doesn't mean that we should accept the criticisms lying down. It certainly is important to provide answers where a suitable occasion exists. It is even more important to make considerable effort to bring about a better understanding of testing on the part of the general public. At the same time, we should not be diverted from the main work we are doing by every blast that may appear in print. I seriously doubt that the criticisms of tests are undermining in any significant way the confidence that those in education, or among the public generally, have in the values of modern objective tests.

If, however, there are times when you are feeling a little discouraged and downhearted because of the belaboring of the critics, let

me give you two quotations from leaders in education who are far better informed on tests than the critics. The first is Dr. Conant:

> However skeptical one may be about the advances claimed by psychologists for their science or however doubtful one may be about the validity of some of the research in education, in the area of tests and measurements even a doubting Thomas must be convinced. The evidence is overwhelming that new procedures and new methods have been developed that, rightly used, can be of great assistance to the teacher and the educational administrator at all levels.

The second quotation is by John Gardner:

> The development of standardized tests is one of the great success stories in the objective study of human behavior. Anyone who understands the problems of mental measurement must be impressed with the technical achievement these instruments represent.

There are, as these quotations imply, many things that are right with testing, and these are best represented by the uses to which tests are put. The amazing thing really is the extent and variety of uses of tests, both in education and in other fields. When one considers that tests have a history of only about fifty years, the part they play in our society is truly amazing. We have by no means reached the ultimate in their usefulness. There are new uses which are being explored, developed, and made operational each year. Parenthetically, I think this increase in the use of tests is the best answer to the critics. If tests were not found to be valuable, their use would decline. Yet all the evidence and the trend lines indicate a steady increase in the use and usefulness of tests.

You are all familiar with the uses of tests, but let me quickly review them. I generally tend to classify them into four categories:

1. The guidance of students with regard to their educational program and future career plans.

2. The placement of students at different levels in particular courses.

3. The evaluation of achievement of the student or the class.

4. The selection of students for the next higher level of education.

In a country with complete decentralization of education, it is particularly important to have such a common basis of comparison as tests provide, since the programs and standards of schools differ so greatly. In fact, one might say that it is only because of tests that a decentralized system such as we have can operate successfully.

These are the four functions that we generally think of in connection with the use of tests in schools and colleges. It is fairly obvious that these are extremely important functions in a democratic society, but they are very general.

Let me be a little more definite and take as examples eight specific uses of tests today:

1. Tests are important in the identification of students of unrecognized talents as, for example, in a case reported by a high school guidance counselor in a large urban school:

> Roger was a quiet boy--the kind who causes no trouble, therefore goes by unnoticed. That was the story of his life--just going along with the group.
>
> The picture changed when Roger took a series of standardized tests in grade 8. His teacher was amazed to find he scored high above grade level in achievement tests and that his IQ was about 136.
>
> Roger has since been encouraged to be more outgoing, and has responded favorably. Today, as a senior, he stands almost at the head of his class; is a class officer; has a five-piece dance band; and plans to enter a seminary in September.
>
> I believe Roger might have gone through unnoticed had not the test results jolted his teacher into realizing that she was dealing with a superior student who was in a shell and needed help to mature.

Critics suggest that there are able students who are overlooked by tests. I venture to say that there are several times as many able students who are discovered by tests.

2. Tests play a vital part in the development of new curricula for the high schools. Each of the curriculum study groups that has set out to develop new programs of study in physics, chemistry, biology, and mathematics has developed tests to go along with its course--in part to provide feedback to those who were developing the courses, in part for the evaluation of the over-all success of the new curricula, and in part to make explicit to teacher and student the objectives of the course.

3. In research on new methods of instruction and the use of new media--like TV, programmed instruction--tests are essential and are regularly used. Without them one would be hard put to it to determine whether or not these new approaches are effective.

4. Likewise, tests are needed in studies of the different efforts that are being made to improve the lot of underprivileged youth, to determine which of the treatments are most effective in giving boys and girls from slum areas the advantages that would lead to fuller development of their capabilities.

5. In a selection of graduate students for National Science Foundation fellowships, tests are essential to a fair and effective selection program.

6. For over ten years now tests have been used as one of the bases for deferring students from the draft, so they might continue and complete their studies before performing their military service.

7. More recently, tests have been used in the first screening of volunteers for the Peace Corps.

8. In the licensing of men and women for professional practice, tests are widely used, as in the case of the National Board of Medical Examiners, and even at the higher level in examinations used by the American Board of Surgery and other specialty boards.

There are other areas of great importance to our society in which tests are beginning to be used and in which their use will certainly increase. One concerns the testing of persons displaced from employment by automation, in order to determine the kinds of jobs for

which they are suited and their capability for further training. A somewhat similar problem concerns those who reach the age of retirement but still have productive years ahead of them. Tests can be helpful, and I am sure will be extensively used in the years ahead, in the guidance of those nearing and past retirement age, so that as longevity and good health extend the years of employment, they will find suitable jobs.

What is right about testing, then, is the important uses to which tests are put. Anyone who canvasses the broad range of test use cannot fail to be impressed by the significant and vital part that tests play in our society.

Behind the uses of tests are the tests themselves, and what is right about testing goes back to the tests themselves. We have today good tests, and many of them--tests that are a far cry from the ones developed in the early years and from the simple-minded tests that appear in magazines. These, unfortunately, are the ones with which the average person is most familiar. Even the critics make their attack on such questions. The reviewer of one of the books critical of testing has pointed out that of the thirty-one questions the critic analyzes, only four have ever appeared in a published test, and sixteen of the items were written by the critic himself! Naturally, this particular critic has a field day.

All of this is not to say that there are not occasional weak, even bad, items in tests, and even in the best tests, but such items are rare. The general level of items is very high.

And tests measure a broader range of qualities than they did thirty or forty years ago. They measure much more than factual knowledge. They measure understanding, the ability to interpret, the ability to apply what is learned in solving new problems, and in some cases they even measure a student's intuitive ability to find the solution to a difficult problem.

It is not only that tests are good, but that we know much more about these tests than we did formerly. Tests are exhaustively studied, and the results of these studies are made available to test users, so that the results may be properly interpreted and effectively used.

In the introductory section of the ETS Annual Report, which will shortly be off the press, there is a rather lengthy statement of what we know about the Scholastic Aptitude Test, and I think it is quite impressive. Over the years a vast number of different studies have been made and much has been learned.

For instance, we know what the effect of intellectual growth on SAT scores is. On the verbal test, the increase in scores attributable to intellectual growth is, on the average, three points a month, or thirty-six points a year.

We also know that the practice effect, or improvement due to having taken the SAT earlier, is about ten points, which, of course, is pretty negligible. Subsequent practice is even less effective.

We know that intensive cramming over a short period of time, six weeks or a term, on which there has been a whole series of studies, has very little, if any, effect in improving a student's scores--certainly not enough to justify the special effort.

We know fatigue and anxiety, which students almost uniformly experience, do not materially affect a student's performance. Even on a test taken at the end of a six-hour day, a student does not do less well than if the test were taken during a half-day session. The effect of anxiety on test performance is more difficult to appraise. It seems likely that there are a few students who are so extremely anxious as to adversely affect their efforts, but on the average anxiety seems to have little effect--and, if anything, it improves performance.

We know, from a whole series of studies, the relationship of the SAT to tests taken at times one, two, or three years earlier. We also know the relationship of the SAT to tests taken four years later at the time of applying to graduate school. The surprising thing is how high the relationship is in all cases. It comes reasonably close to reaching the reliability of the tests themselves.

Finally, we know a great deal about the validity of the SAT as a predictor of college grades in general and in different kinds of institutions. In fact, we know enough about this so that if in a particular study it turns out that the correlation is lower than might be anticipated, one can confidently say that the reason is attributable more to the college program and the grading system than it is to any peculiarity in the SAT itself. We customarily invite colleges that are considering using the SAT to try it out to see whether it predicts success in their institution. Our attitude has been a humble one--that, while the SAT has been found useful in many places, perhaps it may not be as useful at this particular college. My own attitude at the present time is less humble. If a more or less typical correlation is not obtained, the college had better investigate the nature of its program, the accuracy of the grades as given by individual teachers, and the comparability of

grading standards from teacher to teacher. Tests are, in fact, better than the criteria by which they are usually judged.

This kind of information, which is available about the SAT, is also available for many other tests that are widely used at the present time. The goodness of the test, therefore, resides not only in the items which make it up and in the test as a whole, but in the extensive knowledge that we have of the way the test functions.

There is one final matter on which I would like to comment at some length. You have heard the thesis advanced that test items are ambiguous and that bright students are confused by this ambiguity, frequently giving the wrong responses and therefore performing poorly on the test. And this thesis, incidentally, is not unrelated to a concern which has been expressed by college presidents and admissions officers that tests do not discriminate among the very able students--those, for example, above the 600 point on the SAT scale. The evidence, and there is a great deal of it, is all on the side of the effectiveness of tests with high-level, intellectually superior students. Let me cite some of the evidence.

The first study was made by Lindsey Harmon for the National Research Council. He investigated all men who won the Ph. D. in science in the immediate postwar period. He obtained test scores on all of these individuals and equated them to the I. Q. scale. Then he determined the relative yield of Ph. D. 's at different levels of I. Q.:

120 to 130	8 in 1,000
130 to 140	17 in 1,000
140 to 150	30 in 1,000
150 to 160	61 in 1,000
160 to 170	83 in 1,000
over 170	189 in 1,000

This makes an almost perfect progression and clearly indicates that the tests were discriminating just as well at the upper levels as at the lower levels.

A second and related study was made by John Creagar, also for the National Research Council. In this case he studied men who applied for NSF graduate fellowship awards, which, of course, represents a highly able group of graduate students. He divided this group of over 2,000 applicants into five ability groups on the basis of the unweighted

sum of the GRE Aptitude and Advanced test scores. Thirty-five per cent of the men in the lowest ability grouping received a Ph. D., fifty-six per cent in the next lowest, seventy per cent in the middle group, eighty per cent in the second from the top, and eighty-eight per cent in the highest ability group. Lest you think that the award of fellowships may have affected this progression, I should add that among those who were not granted NSF fellowships, the percentage of students in each group who completed the Ph. D. was almost identical, going from thirty-five per cent up to eighty-four per cent.

Since the first two studies relate to scientists, let me turn to a study in the Graduate English Department at Princeton University. A follow-up study was made of students admitted in the years 1952-1955. The Department is a small one, and its standards are very high. Only forty-four students were admitted. The author of the study, a member of the English Department, found that, if a cutting score of 690 had been established on the GRE Verbal Aptitude and Advanced Literature tests, seventeen persons would have been eliminated. Of those seventeen, twelve dropped out during the course of the graduate work, and only five reached the final oral examinations. Of those five, all were rated by the members of the Department as below average. Not one of them was in the group classified as average or above average. There were, in other words, no brilliant students who had failed to do well on the GRE Aptitude and Advanced Literature tests.

The next study was made by Dean Whitla of Harvard. For four classes combined, he studied the relationship of the verbal section of the Scholastic Aptitude Test to success in college, and also the average of the scores on the three Achievement Tests of the College Board with success in college. He made the analysis separately for students from public schools and private schools. The regression lines in all cases are linear right up to the 800 score level. In other words, differences in the 700 to 800 range were just as significant in relation to grades obtained at Harvard as differences in the 500 to 600 range. (See Figure I.)

As a further check on Dean Whitla's finding, we analyzed some records we had of students at Amherst, CalTech and Radcliffe. We studied the relationship of the SAT-V, the SAT-M and the average of the Achievement Test scores in relation to success at each of these institutions. Each of these institutions, of course, is highly selective. In all cases the relationship was essentially linear, with the exception of the verbal section of the SAT at CalTech, which of course is an engineering school and one where the SAT-M would be expected to be more useful. Both it and the average of the Achievement Test discriminate well right up to the top level. At Radcliffe the

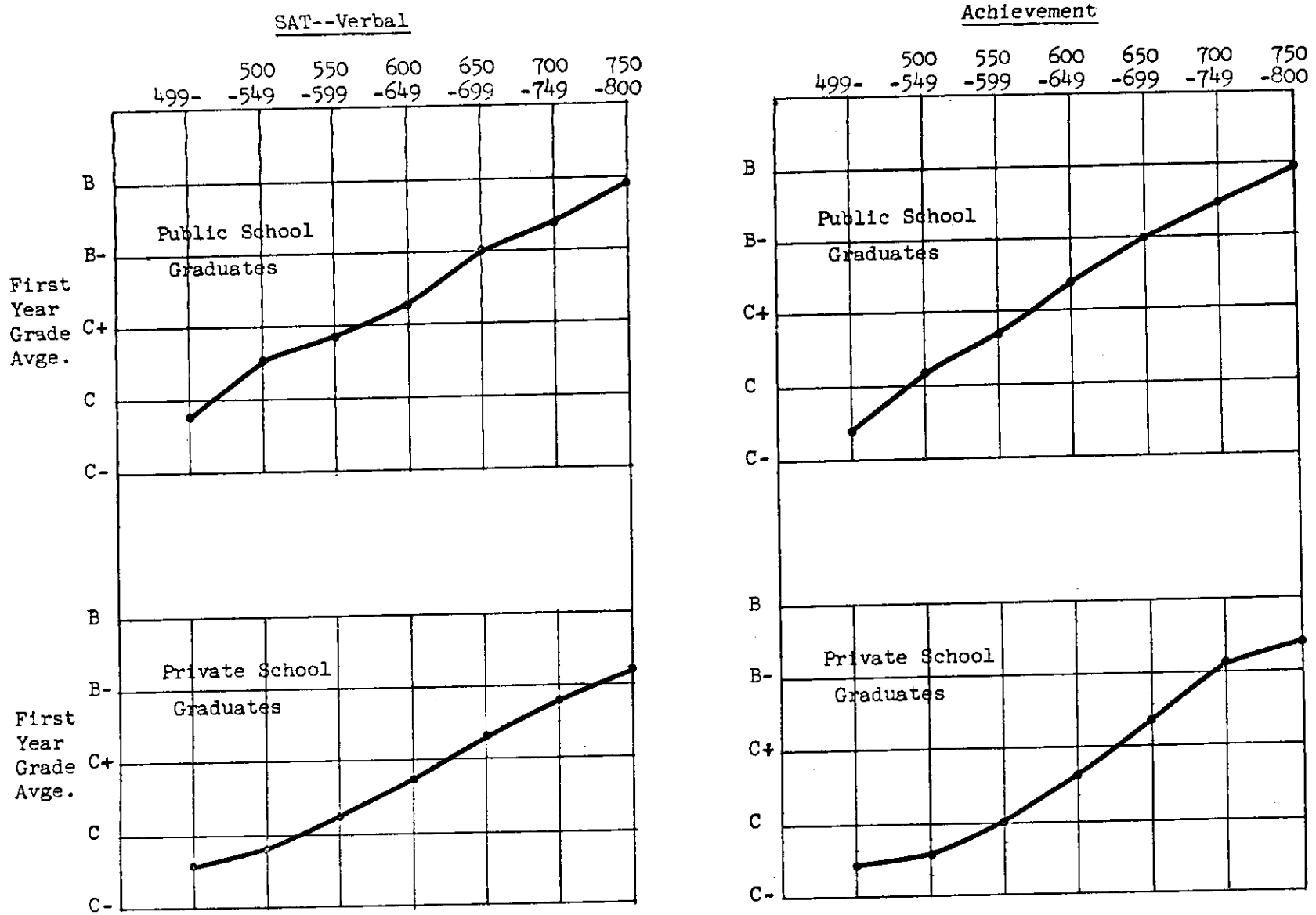

FIGURE I. Profiles of Mean Grades for Harvard Students Grouped by Scores

Source: Study of Classes of 1959, 1960, 1961, 1964 by D. Whitla

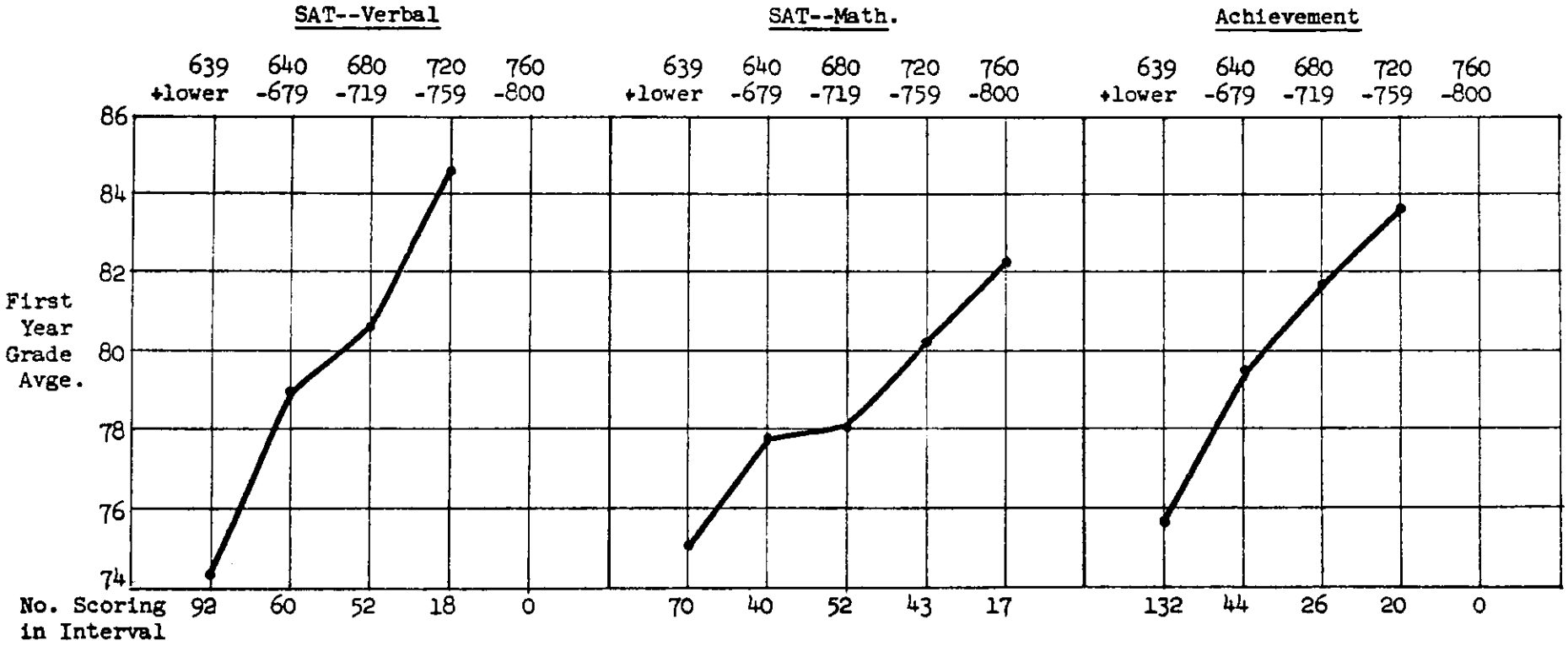

FIGURE II. Profiles of Mean Grades for Students Grouped by Scores

Sample: College A (Male, Liberal Arts)

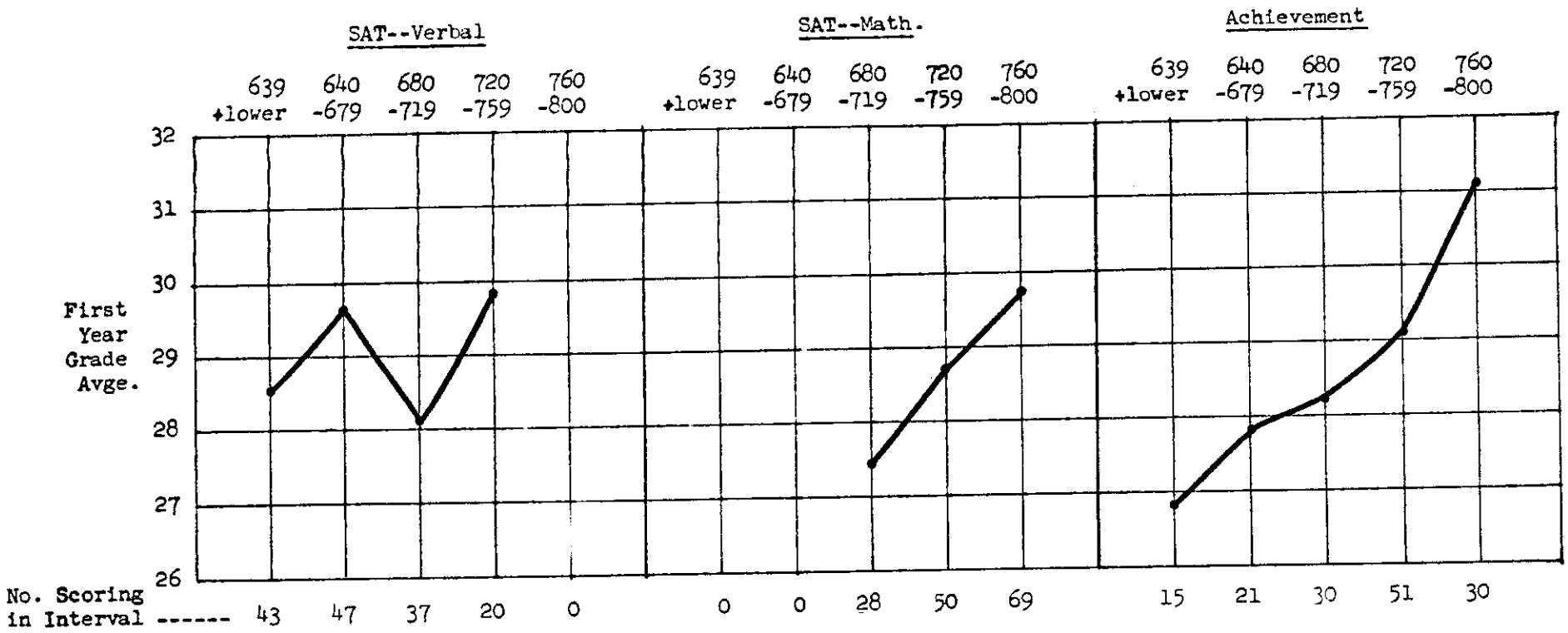

FIGURE III. Profiles of Mean Grades for Students Grouped by Scores

Sample: College B (Male, Engineering and Science)

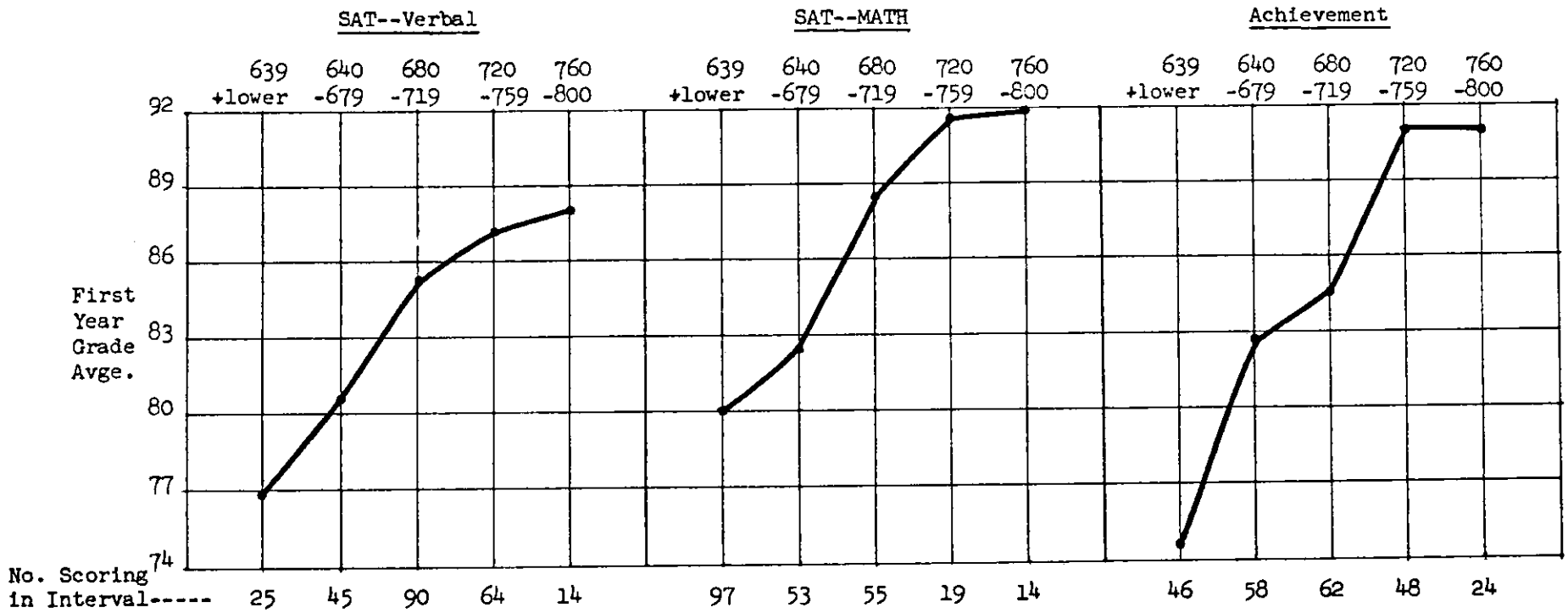

FIGURE IV. Profiles of Mean Grades for Students Grouped by Scores

Sample: College C (Female, Liberal Arts)

Mean First-Year Grades and Per Cent in Top Fifth for College Students Grouped by Test Scores

Test	Score Interval				
	639 and below	640-679	680-719	720-759	760-800
College A (Male, Liberal Arts)					
SAT-V, N*	92	60	52	18	0
Mean Grades	74.3	78.9	80.6	84.5	-
Per Cent in Top Fifth	4.4	16.7	32.7	72.2	-
SAT-M, N	70	40	52	43	17
Mean Grades	75.1	77.8	78.0	80.2	82.2
Per Cent in Top Fifth	5.7	12.5	17.3	41.9	47.0
Achievement, N	132	44	26	20	0
Mean Grades	75.6	79.5	81.6	83.6	-
Per Cent in Top Fifth	6.8	27.3	42.3	60.0	-
College B (Male, Engineering and Scientific)					
SAT-V, N	43	47	37	20	0
Mean Grades	28.5	29.6	28.1	29.8	-
Per Cent in Top Fifth	13.9	29.8	18.9	10.0	-
SAT-M, N	0	0	28	50	69
Mean Grades	-	-	27.4	28.7	29.7
Per Cent in Top Fifth	-	-	10.7	22.0	21.7
Achievement, N	15	21	30	51	30
Mean Grades	26.8	27.8	28.2	29.1	31.1
Per Cent in Top Fifth	6.7	9.5	16.7	19.6	36.7
College C (Female, Liberal Arts)					
SAT-V, N	25	45	90	64	14
Mean Grades	76.9	80.6	85.2	87.1	87.9
Per Cent in Top Fifth	8.0	6.7	20.0	29.7	42.8
SAT-M, N	97	53	55	19	14
Mean Grades	80.0	82.4	88.4	91.6	91.9
Per Cent in Top Fifth	10.3	15.1	27.3	36.8	50.0
Achievement, N	46	58	62	48	24
Mean Grades	74.6	82.8	84.5	91.0	91.0
Per Cent in Top Fifth	8.7	13.8	17.7	31.2	41.7

*Whenever the end intervals contain 10 or fewer cases, those cases were combined with those in the adjacent interval.

regression is linear to 750, and then it levels off. (See Figures II, III and IV and the accompanying table.)

From all these studies, and there are a good many others with the same general purport, one can conclude that tests do not in any systematic way put brilliant students at a disadvantage, at least if we accept as the criterion of brilliance, success in college or in graduate study. The higher the test scores, the greater the probability of success in academic work.

But test scores also relate to success, more generally defined. A study was made some years ago of individuals who had taken the Scholastic Aptitude Test in the early thirties and were later listed in Who's Who or American Men of Science. The results indicated that, at successive levels of test scores, the chances of being listed in Who's Who or American Men of Science went up in the following progression: 1, 2 1/2, 4, 7 and 14, which is a very similar progression to the one that Lindsey Harmon reported in his study of persons obtaining the doctorate.

Now, despite the force of the evidence from these studies, no person professionally trained in the field of testing claims that a test measures all of the important qualities involved in high-level intellectual pursuits, let alone such qualities as motivation, dependability, or independence. Obviously, in attempting to appraise a student, one needs to use other sources of information. This is particularly true with regard to the other qualities that are supportive of the intellectual abilities measured by a scholastic aptitude test. However, this is not to say that the tests are stacked against the brilliant students, nor is it to say that they are less discriminating in the qualities they measure among the most able students.

There is much, then, to be said on the subject of what is right with testing, and I have only scratched the surface. And as the years go by, I am sure that tests will continue to improve, that more of the important qualities will be measured, and more effective use will be made of tests, and that they will play an even more extensive and important role in our society.

All of which means there is a lot of work to be done.

* *