Florida Journal of Educational Research, 1971, 13

INTERPRETING AND USING STANDARDIZED TESTS IN THE 70'S A VIEW FROM ONE MAN'S MOUNTAINTOP

Walter N. Durost

University of New Hampshire

This is a period of upheaval in the schools. Many things are going on to change the character of the schools with respect to school organization and administration as well as the curriculum. We must be prepared to meet the challenge of these changes by providing relevant kinds of information for interpreting test results under all easily imagined conditions. Perhaps it would be helpful to list a few of these changes which gradually are taking place around the country.

Factors Affecting Total Population Norms

1. The ungraded school movement for which Goodlad was the initial champion and which was initally limited to the primary grades is now spreading up through the grades. More and more this is being thought of as a system for individually prescribed curricula which may follow the general pattern or hierarchy which all other students will follow but which does provide the opportunity for the student to move ahead at his own pace.

2. Programmed instruction began to become popular in the very late fifties and early sixties. It essentially was not a new idea but it was thought that we had reached a stage where mechanical or electronic gadgetry might make it more feasible. Even the programmed textbook was considered to be a giant step forward in instruction aiming toward the freeing of the individual from being bound to a class group with whom he was not compatible. This did not prove to gain as wide acceptance as its proponents claimed but it is not a dead horse by any means. Programmed instruction involving slides and cassettes and various other visual aids is very much a part of the curriculum and organization of education at the junior college level especially and there is reason to believe it will persist and have a greater impact at the lower echelons when adequate materials are developed.

3. Computer aided instruction is programmed instruction at its ultimate in many respects. The computer can actually branch from the pupil's wrong response and suggest ways in which he can correct his errors and make progress toward the desired goal. This, however, like programmed instruction imposes a hierarchy which may be simply an artifact growing out of the logical analysis or psychological analysis of the content by the person preparing the program. It often leaves little room for real individual exploration and unfettered thinking. However, it would be unwise for us to ignore the value of computer aided instruction even though the concensus of those who are most deeply concerned with this is that the time is not ripe for it. (See Harvard Review, Fall 1968.)

4. The concept of the open school, which, in a sense, accompanied and grew out of some of the earlier movements to individualize instruction, is gaining ground very rapidly for some rather strange reasons. An open school is one in which the physical environment is such that pupils may move freely from place to place (we can hardly say room to room because there are no rooms; there may be partitions which isolate learning stations so-called.) Under this plan, a student may be working in several different learning stations during the course of the day, depending upon his level of development.

5. Reading, as the central concern of education has been stimulated tremendously by the U.S. Office of Education. Research centers are being established in universities around the country to explore what we know about the teaching of reading and to plan innovative programs so that within a few years *no* individual will leave school without an adequate mastery of essential reading skills.

6. Another new movement threatens to raise a lot of dust, if not more substance, namely, the drive toward criterion reference tests. There is much talk of *criterion reference* tests, testing, or test interpretation. The criterion reference test calls for the complete mastery of an item or skill that is considered to be essential for a subsequent learning in the curriculum. Criterion reference testing seems to have far more relevance to the classroom teacher than it does to nationally derived tests but the movement toward criterion reference testing nationally is marked.

7. In assessing the factors for change in the field of measurement, one cannot overlook THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS, a very viable and potentially powerful influence in the field of curriculum evaluation. Four groups of people, ages 9, 13, 17 and young adults from 26 to 35 have and will subsequently respond to questions in several subject matter areas.

What makes the NATIONAL ASSESSMENT PROGRAM unique is that each person will answer only a few questions but by using many forms within a particular subject matter area many areas of achievement will be assessed within a minimum of pupil time. No child will ever receive a "score" but evidence will be available concerning the proportion of individuals within the tested group able to answer correctly each question in the pool. Judgments must then be made subjectively of the "goodness" or inadequacy of school performance of these age samples.

This all makes for a very complex research pattern and much depends upon the actual randomness of the samples taking each group of questions. This is neither the time nor the place to involve a discussion of NATIONAL ASSESSMENT except to point out that this procedure for obtaining some evidence concerning the achievement of some group with a minimum of testing time in the school has tremendous implications if the sole purpose of testing is restricted to this kind of surveying.

An Evaluation of the Present Status of Achievement and Tests of School Learning Potential (mental ability, intelligence, etc.)

There are literally hundreds, if not thousands, of standardized tests available from all sources but the ones which account for a very substantial portion of all testing done (probably 90% or better) can be counted certainly within the range of a 2-digit number and more probably within the range of a dozen or more tests.

It is obvious in a presentation which is intended to survey a wide range of topics quickly without exploring any of them in depth that I cannot cover, in detail, developments within the field of testing during the past decade. I am going to try, however, to point out what to me seems to be some of the highlights. These will fall generally under four categories, namely,

1. Deviation of test content
2. The writing of items and the refinement of these items by various statistical techniques

3. Standardization of the tests on appropriate populations
4. Provisions for making use of the test information not solely in terms of test score but also in terms of responses to individual items.

I would like now to review what we have accomplished in the 60's in each of these four important areas.

Progress in the 60's

If one surveys the history of testing, it is possible to see certain periods when progress seems to have been extraordinarily rapid in one direction or another. One cannot say the 60's has been one of these periods, but at the same time one must say that the progress has been substantial and perhaps oftentimes overlooked or minimized by those who say that tests in 1969-70 look about the same as they did a half generation ago. In this discussion, I intend to limit myself to a consideration of achievement tests and tests of school learning capacity, omitting any reference to personality measures, interest measures and the like for the sake of brevity.

1. *Test content.* I think it is a fair assessment to say that much greater care is now exercised in the selection of items to go into standardized tests both as regards item types and as regards the importance or essentiality of the information or skill to be measured. The basic source of information concerning the content of an achievement test is still the available instructional material backed up by courses of study produced either at the state or local level or by some federally or nationally sponsored group. The basic source of information still remains the textbook, which is the source giving, in the greatest detail, the day by day content for which the teacher will be responsible. There are few places in the country that have teachers well enough trained to operate independently of the text and in most instances the local course of study is represented by the content of the text, perhaps enriched, or hopefully enriched, by the contribution that the teacher has to make and by reference to other outside supporting information sources.

2. *Writing of test items and refinement of these items by various statistical techniques.* There were notable advances in the 60's in the quality of test item writing but again I am going to restrict myself to consideration of some of the materials prepared, i.e., printed, published, and available, for the current edition of the Metropolitan Achievement Test, c. 1958-9-60.

I, personally, became very much aware of the need of diminishing the effect of guessing, even as early as 1954 and 1955 when we began work on item writing for the Metropolitan, Forms A, B, C, D.

Perhaps the biggest step taken in this direction was the introduction of the "Don't Know" space as a way of providing a student with a place to mark which would convey to the teacher the fact that the information or knowledge or skill called for by the test question was not within his range of achievement. This meant that the teacher could then, by paying attention to the *Don't Know* spaces, identify areas where substantial numbers of pupils had really identified themselves as being in need of additional instruction. Previously we had to depend upon *omits* to provide this information along with some assessment of the percent of responses to the distractors.

Although some informal experimentation was done with the *Don't Know* space prior to publishing the Metropolitan, time did not permit us to carry out substantial controlled experimentation

The most convincing evidence that we have at this moment comes from the item analysis data for the 1970 edition of Metropolitan. The *Don't Know* space *is* used very commonly by students under certain circumstances. It is used far more often by individuals who are at the *low* end of the ability scale than those people at the upper end. Within the subject matter tests, it is used far more frequently by those in the bottom 27% in terms of total number of items answered correctly than by the top 27%. It seems perfectly clear, in terms of available evidence, that this technique does a great deal to reduce guessing simply by changing the psychological impact of the test on the child. Testing becomes a communication situation in which the child is attempting to share both his knowledge and his ignorance with his teacher. Knowledge of what the student does NOT know enables the teacher to teach more effectively subsequent to the test.

Another innovation which was initiated in the 1955-58 Metropolitan revision concerned the spelling and language tests in which the responses to the item itself was a *Right-Wrong* response. In instances like this, the student was required, if he marked the item wrong, to provide, on a space left on the answer sheet, what he considered to be the correct answer. While this did not wholly do away with guessing, our studies indicated that it cut it down very perceptibly. Furthermore, from the point of validity of the Spelling Test, the new item type correlated more highly with the straight dictation than did the multiple choice type, although the margin of superiority was slight. Substantial data concerning this matter are given in the Metropolitan Manual for Interpreting, Page 37. The correlations of the new type item with the dictation type of exercise were generally .90 or better. Unfortunately the use of the writing technique has been dropped in the '70 edition of the Metropolitan because it was messy to deal with on the separate answer sheet. Expediency is often confused with efficiency!

In the 1958 edition of Metropolitan an item type was used in math which called for the actual computation of the right answer on a separate worksheet. This worksheet was to be returned to the teachers with the test answer sheet. (In the lower grades, the test computation was done directly in the test booklet and this was true also for all levels in the standardization tryout.)

The item type consisted of three possible choices, one of which was right unless a 4th choice, "Not Given", was the keyed choice. In all instances, the student also had the option of marking *Don't Know.*

Before the material was tried out in the item analysis edition considerable experiment was carried out on very similar content drawn from the previous edition of Metropolitan. In the first variant of this experiment, the students did out all of their computation first and, subsequently, transferred their answers to the separate answer sheet. It is perfectly clear from the results of this experiment that the task of marking the answer spaces became pretty much a clerical task. *The correlations of the free response and answer sheet scores under these circumstances, were .96 for computation and .98 for problem solving in Grade 5. In Grade 7, the corresponding correlation values were .97 and .99.*

A second experiment was carried out in which the technique

worked out his answer and immediately transferred it to the answer sheet. The results in terms of the correlations between the free response item and the answer sheet item were of similar magnitude.

On the basis of this study, the item was adopted for use in the Metropolitan and separate worksheets were provided for sale by the publisher so that the test booklets would not be consumed.

Unfortunately, I have to report that only a small proportion of the people using the test have purchased the separate arithmetic worksheets. Perhaps they were not publicized enough or perhaps school people were not sufficiently interested in improving the test validity. Just what was done at the local level can only be conjectured but a sample check indicated that the teachers were making use of scrap paper on which the student would do his computation. The very high correlations quoted would not have been obtained unless the student knew that the computation he had done was being turned in so that the teacher could compare his answer sheet response with his test paper response if he wanted to do so. In any case, the item is one which is machine scoreable or optically scannable and yet the basic validity of the item has been protected by requiring the individual to do the task he was supposed to do. In the item analysis and normative editions of the 1970 series of Metropolitan math tests, expendable booklets were used but no provision is made to sell separate math work sheets.

In the mid-60's, construction of a new measure of mental ability called the Analysis of Learning Potential was begun. This series is now available. These tests, ranging from Kindergarten through senior high school, (pre-kindergarten and college forms are potentially available also) incorporate many of the suggestions mentioned above. In addition, the tryout tests for the Analysis of Learning Potential were selected only after a psychological analysis of the component skills needed for success in each particular grade was done. For every test tried out, (of which there were many more than were finally used) a "justification" sheet was provided by the author which gave his rationale for the test. These were circulated and discussed by the authors and the Harcourt staff before the test was put "in the mill."

The Analysis of Learning Potential makes no assumption that it measures something called "intelligence" which has an

in and of itself. Instead, the only claim made is that we are measuring mental abilities which are of a non-school variety which correlate highly with measures of specific in-school achievement. Many new item types were tried out. Those which survived the test of practicality will be found in the presently published tests.

Problems in the Establishment of National Norms

The techniques and procedures used to establish national norms for standardized tests have developed tremendously over the years since the Otis Group Test of Intelligence or the Stanford Achievement Test were first standardized. During this period, many people have attacked the whole concept of national norms as being impractical and essentially meaningless because of the uncritical combination of populations differing from one another in basic ways either directly or indirectly related to scores on either intelligence or achievement tests. These people have advocated instead the use of local norms as being preferable. A very good case can be made for local norms for certain purposes, although not to the exclusion of the national norms.

Different communities do differ widely in their socioeconomic backgrounds, ethnic composition, and even differ in ways associated with subtle climatic factors. Over and beyond this, there are differences in every phase of the curriculum arising from fundamental differences in attitudes and value systems inextricably mixed up with some of the socioeconomic and other factors mentioned above. To combine samples of community performance on standardized tests without regard to these many peripheral influences has long been recognized as being less than satisfactory. This unspecified "mix," largely due to chance, is perhaps the major reason why test norms vary so much from one standardized test battery to another.

It must be obvious that the publisher who fails to provide national norms on any test to be used all over the country would be courting disaster. The alternative, therefore, has been to move steadily in the direction of refining the normative procedure to obtain norms which would reflect in proper proportion the contribution of the various impinging factors known to affect psychological and educational test outcomes.

This refinement of procedure has not been one that has moved as rapidly from the shotgun approach of the 1930's to the more refined technique of the late '60's and '70's. It is an historical fact that members of the staff of World Book Company, and

subsequently of Harcourt, have made notable contributions to changes in this area, beginning with a study by Roger T. Lennon in the early '50's and continuing through a series of both formal and informal investigations based upon the experience of World Book Company and Harcourt in carrying out such standardizations. The most recently completed study in this area is that of Thomas Hogan which is entitled, *Socioeconomic Community Variables as Predictors of Test Performance.*

The net result of these studies has been the development of a procedure for stratified random sampling which takes into account the socioeconomic factors, regional differences, etc., which seem to affect test results. This new approach to the establishment of national norms, contrasting with the approach characterized by the 1958 edition of Metropolitan and the 1964 edition of Stanford, where only casual attention was paid to these factors, was first implemented in the standardization of the Otis-Lennon Tests of Mental Ability. The writeup of the procedure used in standardizing this test as it appears in the Manual for Administration will undoubtedly represent one bench mark in the development of a more systematic approach to test interpretation. Along with this, but not directly related to it, has been the general adoption of the idea of having comprehensive scaled scores covering the entire range of grades for which a particular battery of tests is intended. World Book Company's first major attempt to do this type of cross-grade scaling was in connection with the 1945 edition of the Metropolitan Achievement Test in which a simplified Thurstone approach was used. In retrospect one can say that the abandonment of this technique in the Metro 1958 edition was a mistake, although this abandonment was largely due to the conclusion that the cross-battery scaling had not had the practical results that had been anticipated. Suffice it to say that this procedure now seems to be well established in that it has been used, not only with the Otis-Lennon, but with the Analysis of Learning Potential and more recently with Metro 70.

To return to the main issue, namely, standardization of tests, the next major attempt to apply the stratified random sample approach, following closely the pattern of the Otis-Lennon, was the standardization of the new test of learning potential called the Analysis of Learning Potential. This test has been described briefly elsewhere and no further comment is needed here.

Finally, the procedure of stratified random sampling has been most extensively applied in the achievement area in the recent standardization of Metro 70 in which refinements have been adopted based largely on the studies by Hogan.

It is also significant that the Otis-Lennon test, not only was independently standardized but has been used also in connection with the revision of the Metropolitan, which in effect amounted to a cross-validation of the Otis-Lennon norms. Since the Analysis of Learning Potential and the Otis-Lennon also have been equated to each other, a tremendous data base has been established for future use in obtaining national norms for either intelligence or achievement tests and for either age-based or grade-based test norms.

The care with which these national norms populations have been selected to secure representativeness now lets us make some generalizations about the true shape of the distribution of scores for various tests. The outcome is most interesting. In a very wide variety of test materials constructed independently of what might be the outcome in terms of the distribution of scores on a national sample, we have nevertheless obtained essentially normal distributions. These distributions have been further characterized by a tremendously wide distribution of scores for any unselected age or grade group, from almost a perfect score to a near zero score for any systematically defined strata of the population. This fact in and of itself is not particularly surprising because it is obvious that the factors impinging on the test score are so multitudinous that the conditions needed to satisfy the requirements of a normal curve, or the curve of chance, are clearly met. While this is more evident in the case of mental ability tests which are not directly school oriented, the statement applies almost as adequately to tests in the field of specific in-school instruction or achievement. As a result of all of these studies it now seems reasonable to expect a normal distribution of scores on any test which has a sufficient range of difficulty to spread out the existing talent. This fact, in itself, makes the application of normalized scaling techniques sensible and appropriate.

This discussion to this point fails to face up to the most vital question, namely, are national norms applicable within a local situation? I think the answer to this is that such national norms do represent the performance of a large and carefully defined population of individuals all over the country and because

of this fact, they do provide a backdrop or yardstick against which the individual community may compare itself especially where averages are concerned. One would not, however, expect that the distribution of scores which is characteristic of a national population would similarly apply at the local level. To put this differently, the variability of scores at the local level is more likely to be less than it is at the national level, rather than the reverse. The comparison of local means, however, (whether "local" means a community, a county, or state, or some other large unit is immaterial) is made meaningful and helpful by the confidence generated in the national norms due to the care with which they have been obtained.

Perhaps the next point to be made is one that is rather subtle and difficult to communicate without a substantial amount of sophistication in this area on the part of the reader. One can randomize such factors as teachers' salaries, length of teaching experience, amount of dollars spent for education per pupil, the educational level of parents, etc., because all of these factors are peripheral. It is not possible to stratify before the fact on some test of learning potential, such as the Otis-Lennon, and yet such data do constitute a tremendously important source of reassurance that the norms are representative of the total population on mental ability.

No local community is justified in interpreting its achievement outcomes without also considering its status on some such test of learning ability. There are no hard data to show that high achievement can be secured with a low level of measured mental ability when the nature of instruction is also typical. In other words, learning ability tests exist primarily as a way of assessing the potential for school learning on the part of the students being taught and this applies to whole communities as well as to individuals.

In recent years and months, there have been powerful forces at work which are striving to do away with "intelligence tests" as being unfair to certain ethnic groups. Assuming now that we grant immediately that the intelligence tests so-called are not really measures of inherited mental ability, an unbiased unemotional consideration of all the factors involved should result in a change of heart on the part of these people with respect to the use of such tests because of the contribution that they make in indicating the "readiness" of any population to move into any area of school learning. There are two important factors involved:

1. Most of the better tests labeled as intelligence, mental ability, learning ability, learning potential, and the like, strive hard to measure learning which comes about due to factors in the environment more or less available to all children, as contrasted to achievement tests which attempt to measure outcomes related directly to instruction in school. This endeavor to generalize the items for mental abilities tests to free them from specific dependence upon in-school instruction is not something that is accomplished by statistical analysis but rather it is a matter of logic. The items constructed must be examined with great care to see that they are not unduly influenced by biasing factors existing either in the schools or in the general environment. The more recently constructed tests of this nature have done a fairly good job of achieving such a goal, largely due to the fact that they are examined by many people of widely differing backgrounds at the item development state who search carefully for bias and for ambiguity which might affect the validity of the outcome. This approach has been most systematically carried out for the Analysis of Learning Potential and Metro '70.

2. The next most important consideration for the person in charge of testing in any local community must be the search for tests of learning potential which have norms comparable to the norms available for the achievement measure he intends to administer. Most of the publishers have moved strongly in the direction of providing this kind of pairing of ability and achievement tests. It is particularly noteworthy, however, in the case of the Otis-Lennon and the Analysis of Learning Potential where tremendous amounts of national data have been accumulated to assure comparability of results. Probably the most carefully conceived program for doing this is the one recently carried out with Metro '70 in which the Otis-Lennon tests were an integral part of the standardization procedure and the norm populations are clearly specified as to the learning potential parameters measured by this test. Along with the Otis-Lennon, work now in progress and shortly to be finished will similarly tie the Analysis of Learning Potential to Metro '70 and to Stanford. We thus see that we are now at the point of reaching a goal long sought after under one guise or another, namely the creation of a stable, constant, normative population which can be re-created by anyone who wishes to invest the time and expense to do it. Over the years this concept has been called by various names, such as "the standard million" concept or the anchor test concept.

The Modal Age or Age-Controlled Norm Population

In the work described above, nothing has been said about the additional requirement of good norms for use in interpreting individual test scores that the population shall be a constant one across grade levels from the beginning of school to the end in the sense that systematic bias is not introduced by administrative policies regarding entrance age, promotion, and retardation. For such norms to be comparable, it is necessary to establish the fact that both the age range of the population and the level of exposure to instruction have been controlled. It is easiest to see this concept in the light of the present grade structure, although the concept itself is certainly not tied to gradeness. In May, 1940, Truman L. Kelley wrote an article in the Harvard Educational Review entitled "Ridge Route Norms." In this article Kelley attacked this problem of lack of comparability in age norms versus grade norms and suggested basing norms upon a modal group at each grade to be obtained by a rather complex statistical procedure. As the Director of the Test Division at World Book Company at that time and the person administratively responsible for standardizing the Stanford Achievement Test 1942 edition, I conferred with the authors of the Stanford Test with the idea of applying this technique to the norming of these tests and it was agreed that we should do so. However, a simplified procedure was used to obtain this modal age group. A distribution of ages was made on a one month basis and that range of 12 months of age showing the largest number of cases for any similar 12 month range was identified and called the modal age group. Subsequent cross-validation showed that this simple procedure resulted in almost the same population being identified that would have been chosen by Kelley's more sophisticated approach.

The modal age groups thus identified from grade to grade were found to move up exactly one year for each subsequent year of grade. It was also found that this modal age group did result in the selection of individuals who were somewhat superior to the total grade population in mental ability as tested by available mental ability tests. In spite of this fact, the Stanford authors provided norms for this edition of this series based upon modal age groups.

At this point one is forced to recognize the inevitable fact that any innovation of this sort, especially if it is not completely understood and does upset conventional procedures for tests interpretation, has a hard job of being accepted generally within

the short range. The 1940 Stanford norms were objected to because they were "too hard" especially when averages of total grade populations were compared with the published norms.

In the 1958 edition of Metropolitan the single 12 month modal year concept was broadened to include 18 months of age to allow for differences in policy with respect to age of entrance and promotion, and the norms for this edition of Metropolitan were based upon this new modal age concept which was renamed the "age control sample." Perhaps in part because Metropolitan was standardized in the fall and, in larger measure, due to the increase in the age range represented at each grade, Metropolitan norms have been widely accepted without question, although certain large cities, especially, have preferred to use total grade population norms.

The question has often arisen as to what should be done at the local level about identifying the *local* age control sample and making comparisons with national norms on the basis of the performance of this group rather than on the total grade population. This procedure, undoubtedly, is a more precise way of making such comparisons and is highly recommended, but since this involves a dual analysis, namely, for the age control sample versus the total population, it has rarely been done.

The 1964 edition of Stanford abandoned the modal age concept and for all practical purposes reverted to the total population base.

In Metro '70 total population norms have been provided in order to meet the demands of communities wanting to compare their total populations with a national total population-type norm. However, norms based on the age-controlled population also will be provided and users will be urged to use these norms, especially when making longitudinal studies of individuals where comparison with a population of more carefully controlled composition becomes essential if year by year results are to be considered comparable.

The interpretation of results for individual students is, after all, the major purpose of testing. Tests of the length and complexity of those used in our present survey tests were not necessary in order to establish community averages, but there is a serious question as to whether such attempts to get "quickie" community averages is worth the time and the effort it takes when the expenditure of a small amount of additional testing time will yield data which has great relevance to instruction, guidance, as well as administration.

Care in establishing the representative character of the norm population is only the first step to a more valid interpretation of test results. There remains the task of choosing the methods of data analysis that will be meaningful to the user in a variety of situations. To some extent the publisher as well as the user is the captive of precedent and there is need for some clear thinking and courageous action to correct some of the present inadequacies.

There are two general types of norms which must be considered. These are the "trend line" norms and the "peer group" norms. Grade equivalents and age equivalents represent the first type; percentile ranks and various types of standard scores the second.

Grade equivalents are very widely used – or perhaps I should say misused. Grade equivalents and age equivalents are based upon a simple and basically sound idea. The best way to evaluate an individual's performance is to compare it to the average of some group. Thus, the score that conforms to the average of the norm group at some defined level of progress (either age or grade) is named accordingly. For example, the score that is just average for a defined group of children 10 years and 5 months of age is said to have an age equivalent of 10-5; a score that exactly conforms to the average score earned during the second month of Grade 5 (usually October) has a grade equivalent of 5.2. All would be well if the matter stopped here but in practice it does not.

Age equivalents have almost disappeared from the scene, although they may be due for a revival if the ungraded school becomes widely accepted. Consequently, the age norm will not be discussed here.

Let us concentrate attention on grade equivalents and see why they are now in disrepute among many of the professionals in test construction and utilization.

The procedure of establishing grade equivalents is essentially simple but fallacious in serious ways. The steps are as follows:

1. Scores from adjacent batteries or levels are converted to some kind of continuous score scale. For present purposes, let us assume that this is a scaled score of the Thurstone type or some other type of score that will permit comparison of the norm lines from test to test. The simplest method is to use the mid-battery raw scores for the purpose.

2. The averages of successive grades (usually the median) is plotted against the time of year of testing. Metro '70 Fall norms are based upon tests given in October; therefore the average at each grade gets a grade equivalent of x.2, x representing any grade. A ten-month school year is assumed.

3. A smooth line is drawn through the plotted points and intermediate scores are assigned grade equivalents by interpolation. Herein lie the most damaging fallacies. The school year is NOT typically 10 months long. Nine months would be a better guess. The plotted points, moreover, are 12 months apart; not even 10. To divide the gain from x.2 of one grade to x.2 of another into ten parts is just plain wrong unless ALL learning in a subject takes place within the school environment and nothing at all is learned during the vacation months. Some tests, notably reading and vocabulary, show as much gain during the summer as during the school year; others, such as arithmetic computation, show little progress during the "off" season.

The growing practice of evaluating special programs such as Title I projects in remedial reading or the outcomes of special instruction under contract arrangements in terms of grade equivalents, is patently ridiculous unless the gains reflect what takes place during a normal school year or more particularly the period of time between pre- and post-testing which is more usually 7 calendar months.

Consider the pre-post test situation in a single test such as reading. What would be the expected or normal gain over seven months of school? In the Stanford Elementary Reading Test the raw score gain from 3.2 to 4.2 is 8 points. If divided by ten (tenths of a school year or, more precisely, months of GRADE) a seven month span between tests would mean an expected gain of 5.6 points of score. If the 8 raw score points is divided by 12 months of the calendar year the expected gain would be two-thirds of a point of score per calendar month or 4.7 points of score over the seven-month teaching period.

Since no one knows exactly what the rate of gain is over seven months, we felt it necessary, in New Hampshire, to test a random sample of the state school population at four grade levels to provide a more reasonable estimate of expected growth against which to evaluate our Title I reading program. Metro 70, let it be noted, will have both Fall and Spring norms.

4. The smoothed norm line, drawn through the plotted points from lowest to highest tested grades next is EXTRAPOLATED to assign grade equalivants to a large number of points of score falling outside the range bounded by the lowest and highest plotted medians. This procedure is sheer guess-timation. That does not, in itself, condemn it. The lines are usually so smooth that the extensions are not hard to make *if one follows the trend of the line.* However, 1.0 is the lowest possible grade value under this system and many scores will normally be left without assigned grade equivalents if the test is well designed to measure existing ranges of reading ability even in Grade 2. At the upper end of the curve the results of extrapolation are fantastic. Grade equivalents have been widely given up to 12.0 although by no stretch of the imagination can any segment of the elementary school curriculum be considered as being extended in fact beyond an absolute limit of 8.9 or 9.0. For 50 percent or more of the students tested, the earned scores at Grade 8 thus are assigned totally fictitious grade equivalent values.

Even the extrapolation of the middle battery scores to upper and higher grade equivalents than the grades tested is fallacious since the content of the test almost never represents the curriculum content of grades represented by these extreme grade equivalents.

Grade equivalents have only one legitimate use, namely, to evaluate the extent to which average scores of grade groups deviate from the norm. Even this use may be misleading for extremely high and low ability groups.

To summarize, grade equivalents in the hands of unsophisticated persons almost always speak with a forked tongue.

1. They are not comparable from grade to grade.

2. They are not comparable from test to test.

3. They are nearly meaningless when extrapolated beyond the limits of the plotted points.

Peer group norms (percentile ranks, and standard scores of various kinds) provide the best basis for determining the exceptionality of any student's measured performance. Even the Deviation IQ, now almost universally used, is a peer group standard score. Its constancy (approximate for individuals) derives from the fact that the actual score distributions are, in effect, re-scaled at each significantly different age level to counteract increase in variability associated with age.

In many ways the percentile rank norm is the most "face valid" type of peer group norm. It can be thought of simply as the individual's rank in a representative group of 100 individuals or as his place in a cumulative percentage distribution. Percentile ranks cannot justifiably be treated statistically. The units are not separated by equal differences in achievement if the score distribution is bell-shaped. They can be equal only in a perfectly rectangular distribution which never is found for any typically constructed test. For example, percentile ranks cannot be used to measure gains. Percentile rank bands, widely advocated as an antidote for over-precise interpretation of test scores in view of measurement error, certainly have the virtue of preventing any such statistical manipulation!

Of the various standard scores used to interpret test scores, the *stanine* is now by far the most widely used and generally the most satisfactory. The stanine was developed during the war by the Air Crew Selection Program under John Flanagan. I have talked to Dr. Flanagan about this matter and was told that it would be impossible to say what individual was responsible for first suggesting the technique or for giving the resulting 9-step score a name. Apparently, the idea came out of a group discussion of the necessity of finding a single digit score in order to reduce the amount of work involved in relating some 30 or 40 different tests used in the selection of candidates for such Air Force positions as bombardier, navigator, pilot, etc.

For two or three years after World War II this concept languished. However, I had worked with these scores at Air Service Command and in the Adjutant General's office and had found them so useful in dealing with adults that after I returned to my position as Head of the Test Department at World Book Company, I began to experiment with the use of stanines for interpreting pupil's test scores on standardized tests. I extended my experimentation in this field greatly after I went to Boston University and, after I left Boston University to establish my own Test Service and Advisement Center in New Hampshire, I adopted it as the basic method of test score interpretation.

By 1950 most major World Book Company tests were provided with stanines as a standard part of the interpretative machinery. This practice is now universal. Evidence is available that indicates that stanines now are the primary basis for interpreting scores in nearly all of the larger cities in the United States and in many many small communities around the country. Stanines are also provided routinely as part of Harcourt's scoring service.

Advantages of the stanine units. Stanines, being normalized standard scores with a unit of 1/2 a standard deviation centering around the mean (mean of 5, SD of 2), provide a basis for comparing scores either from test to test, or from pupil to pupil so long as the population on which the stanines is based is the same. Stanines from grade to grade also are very generally comparable if there is no violent change in the composition of the grade group. Since stanines involve an area transformation all that is required to establish them is to arrange the scores or numbers to be transformed into rank order and to lay off the standard percents of cases for each level. Thus, they are extremely easy to compute, especially if any kind of computing equipment is available, from a desk calculator on up to a computer, which will give cumulative percentages. (In point of fact, these cumulative percentages are in themselves the basis of the oft-desired percentile ranks.)

Stanine scores, since they are comparable in their variability from test to test when based on the same population, can be combined in various ways to obtain composite scores of various kinds. The composite prognostic score with which I experimented extensively in Pinellas County while I was the Director of Educational Services Division has proved to be a very effective way of getting an overall measure of the individual's school learning potential on the basis of a combination of capacity and achievement measures. Such composite scores were highly predictive of college success in the Freshman class of Florida institutions of higher learning as reported by June Hopper.*

I have recently been working with the State Department of Education in New Hampshire as a Consultant on their testing program which is being carried out under the auspices of Title I. For this program, and for previous statewide testing programs going back to 1952, stanines and composite prognostic scores have been provided. In one longitudinal study of composite prognostic scores done in Concord, New Hampshire, the composite prognostic scores computed on the basis of data gathered as long ago as 1957 have been compared over two-year periods and the composites for Grades 4, 6, 8 and 10 were compared with rank-in-class at high school graduation as an ultimate criterion.

*Test Service Bulletin No. 101, A Prognostic Score to Predict Senior Placement Test Performance. Harcourt Brace Janovich, Inc.

The amazing result is that the composite prognostic score at Grade 4 alone predicts rank-in-class at the end of the 12th grade fairly well in spite of all of the difficulties involved, the correlation being .68; at Grade 6 the correlation is .73 and at Grade 8 it is .76. The class ranks were transformed to stanines for this study, a very simple procedure.

I have tried to approach the interpretation of stanine bivariate charts in a different way by setting off a 3-step wide band diagonally across a bivariate chart running from lower left to upper right when the predictor variable is on the vertical axis and the criterion is on the horizontal axis. When the correlation is in the order .68, as it is for Grade 4 composite versus rank-in-class, we find that 74% of the children will fall within this 3-step corridor. By using a 3-step corridor in this way, we allow for the standard error of measurement which in almost every instance is less than 1 stanine, if the test is reasonably reliable. One can then say that the cases falling above or below this corridor are ones where the child is performing in a manner inconsistent with his measured potential. Only 15% have measured potential higher than rank-in-class connected to a stanine; similarly 11% have measured potential lower than rank-in-class (stanine.)

In a similar way, it is possible to analyze a child's progress in reading or arithmetic or any other subject matter field. In any systematic testing program, different tests will be used which may measure different components of reading or different components of arithmetic during any 10-year period. Almost inevitably the curriculum will change over a decade. This was true in Concord, New Hampshire where this study was made. Yet the relation between subsequent composite prognostic scores remains high. The consistency of performance really is amazing. One rarely finds an individual who starts off with a high composite prognostic score in Grade 4 (Stanine 7, 8, or 9) who reverts in any subsequent measure to a position substantially below average. When this actually occurs, case study investigations have repeatedly shown adequate reasons for the drop or increase. Longitudinal studies, such as the Concord study, should also include some way of combining measures from year to year to get a cumulative index The most appropriate way of doing this remains to be spelled out.

This high consistency of the prognostic score is somewhat disturbing to some school personnel in that it suggests that a child's rate of learning is a characteristic not easily modified.

Better methods of instruction may only raise the entire distribution rather than raise it more for the less able than for the more able. Only the child with a clearly remedial defect in reading or math will show any substantial change in prognostic score after corrective instruction.

"The Pygmalion in the Classroom" Phenomenon, Fact or Fancy?

Not too long ago a book was published which received wide attention from the general public as well as among educators. It was entitled, *Pygmalion in the Classroom* bv Robert Rosenthal and Lenore Jacobson. This book sets forth the notion that testing (or by implication any other systematic objective data gathering effort) where the outcomes are made known to the teacher, will result in the teacher establishing a mental attitude toward the child which will predetermine his success or failure. Most startlingly, the claim was made that if the teacher was led to believe a child had a high IQ, that child would achieve in proportion to his concept of him whereas children thought to have a low IQ would consistently fail.

Let's first consider whether or not the Pygmalion idea is, in fact, true. The research study itself on the basis of which the authors of *Pygmalion in the Classroom* make their claims has been shot full of holes as being a shoddy, poorly designed. badly carried out piece of work. It has even been shown that many of the teachers involved in this so-called experiment actually put the test data away in a desk drawer and never looked at them from the beginning to the end of the experiment.

I think it must be granted that every teacher is bound to have some concept of what a particular child is like, of what his learning potential is, what his social and adaptive behavior is, what his level of aspiration is, etc., etc. Few teachers are objective enough to take into account the fact that many of these phenomenon arise from their own attitudes toward children, which in turn arise from totally irrelevant factors such as the skin color of the child or the kind of socio-economic background from which he comes. Test data, at least, reveal something objective about the child and while standardized tests of mental ability or learning capacity certainly do not measure inherited intelligence solely, they, nevertheless, do indicate the existing level of functioning in the school setting of the child at any particular moment. The

composite prognostic score, however, depends more on what the child has learned in and out of school in the basic skills of reading and math than on measured mental ability and this is a more functional approach to determining actual readiness or learning potential at any given point in time.

One follow-up study of the composite prognostic score was done here in Florida, making use of data from Pinellas County. In this study, published as Test Service Bulletin No. 101 by Harcourt, Brace & World, June Hopper, author, it was found that success in college was startlingly well predicted by the composite prognostic scores obtained on the basis of *ninth* grade testing alone. Correlation of the math-science composite prognostic score with total standing, 12th grade statewide test, was .87 in 1961.

The qualifying score on the statewide 12th grade test was 300. This corresponds to a statewide percentile rank of 50 or a stanine of 5.

About 11% of all the students tested in Grade 9 in 1957-58 and retested in 1961 with the college placement test earned stanines of 5 or less in 57-58 but earned scores of 300 or *more* on the state test in 1961. Of ***all*** students in the 5 stanine or below category, 19% earned scores of 300 or more on the 12th grade test.

It is said that the average score on the 12th grade test for students doing "C" work in the universities was 350. Only 9% of those earning composite stanines of 5 or lower reached or exceeded 350 on the 12th grade program. This study also concludes that only three out of four with stanines of 7 or better at the 9th grade could be expected to do average or "C" work in college.

What Will (Should) Happen in the Field of Evaluation During the 70's

It has been necessary to cover some of the developments of the 60's and the status of testing and evaluation at the present moment in order to lay a groundwork for some of the things which I personally feel are going to happen in the 70's. Some of my predictions (if they might be called that) make very basic assumptions which, if not true, can greatly change the predicted outcome. Nevertheless, I will have succeeded in my purpose if I can arouse some interest in possible future developments.

I have been concerned for 35 years with the lack of adequate pre-service training to prepare teachers to undertake this very essential part of the total educational process, nor has this lack been compensated for by adequate in-service training. Sometimes it has been said that the ideal teaching situation is a great teacher on one end of a log and a student on the other. If so, it is because of the dialogue that takes place between the two, and this dialogue, by its very nature, constitutes perhaps the most efficient way of evaluating the performance and learning of the individual student.

Due to the number of children we have to take care of and the per-pupil cost of education under present circumstances, anything approaching this kind of an individual dialogue is hard to imagine. However, the move toward greater individualization of education which is now beginning and spreading more and more rapidly will eventually dominate the educational system in this country. My enthusiasm for this point of view arises from the fact that the present system of lumping children together is so completely in violation of everything we know about conditions of learning.

Naturally, if this individualization does take place a whole new area of development is opened up for the use of all kinds of measuring and evaluating devices including tests. Their use no longer would be optional but would become mandatory because no one would otherwise know when a child was really ready to move along to the next stage in his learning development.

This is not, however, to be interpreted as a plea for standardization of content in any subject field in the form of some hard-and-fast hierarchy especially in the less-structured subjects such as literature, social science, etc. A stultified hierarchy of social-studies learning or science learning could destroy much of the spontaneity that should be characteristic of the acquisition of knowledge in these rapidly changing areas. On the other hand, a certain degree of emphasis on an hierarchy within the basic skills may be desirable, especially in the lower grades and, of course, is easily measurable. Hierarchy or not, *there must be stated and measurable goals of instruction and tests must be made to test them.*

The present procedure of establishing national norms on carefully selected stratified random samples may persist for some time—perhaps for more than one lifetime. However, there will

be more and more dependence upon local norms as a supplement to national norms. Where national norms are used to provide some standard basis of reference, an attempt will be made to define the norm population ever more precisely from grade to grade or from developmental stage to developmental stage. Retardation in its present form will, of course, become nonexistent when a child is allowed to proceed at his own pace as in an ungraded school. Similarly, we don't need to concern ourselves with acceleration as an administrative device since the ablest pupils also will be moving along at their pace, either in terms of mastering more and more difficult content or in broadening their understanding of content normally taught in the elementary and early secondary grades.

There has been much talk of late about the present system of evaluation in the schools which condemns a large portion of children to the stigma of failure because they do not maintain their position in something artifical we call a grade. This stigma of failure will disappear in a system which is completely individualized, but there will still remain the constant task on the part of the teacher and the parents of children to insure that a child is working up to his capacity. This very need, it seems to me, guarantees the continued existence of measures of school learning potential by whatever name we choose to call them.

It would be a fortunate thing indeed if we could completely remove from our testing literature the term "intelligence" testing or even "mental ability" testing and substitute for it something like "learning potential" as has been done in the case of the Analysis of Learning Potential. Such measures will recognize the fact that they are valid only as a measure of the complex of mental traits necessary for success at a particular level in school, recognizing that these mental traits may not indeed be the ones which would be most predictive of success in other situations. Measures of intellectual power (basically the mental age idea brought up to date) and measures of brightness or exceptionality where age is held constant are both very necessary, but, of the two, the measures of power are better for relating capacity to achieve to actual achievement, since achievement tests are usually power measures. Hence, we may see a tendency to group children for analysis purposes (but not for instruction) into groups that are more nearly homogeneous with respect to their capacity for learning. Furthermore, the time

will come when we will investigate these parameters independently for boys and girls and will expect and foster differences in the rate of learning within different subject matter areas which are related to the strengths and weaknesses associated with being a boy or a girl.

Perhaps before I continue to dream in such broad and comprehensive terms I should bring myself back to a consideration of some of the mechanics of testing in which we will see changes in the next two or three possibly and certainly within the next five years.

1. Item analysis will assume much greater importance in the school of the future. We will recognize that the test item is the touchstone or basic building stone for a test. In fact, really convincing arguments can be made for restricting the combining of items into some kind of score only in those areas where the tasks have some definable homogeneity, which often is not the case in our present tests. This is especially true in the subject matter fields such as science and social studies.

In the course of improving our test items, we will steadily move away from items subject to guessing and we will make use of techniques such as the "Don't Know" space to make the items reflect the actual knowledge of the child more precisely. The time may come even when we can dispense with the multiple choice question altogether and use equipment for test processing which will scan optically a freely written choice and evaluate its correctness or incorrectness. I don't see this as being immediately on the horizon or even as a highly essential development at this time. I do see it as being very important that we make the test item our servant and not our master. In other words, we must not allow the dictates of the optical scanner or the computer to tell us what kind of test items we can have if this means that the test items are intrinstically less valid than they would be otherwise.

Actual performance of a national sample of children on selected items in areas such as arithmetic is sadly out of line with the "expected" performance if one may judge by the level at which a topic is introduced.

The concept of criterion reference testing is merely a dream until we reach the point where we can say the total experience of the child has been such that he should, by a given stage in his development, have mastered the knowledge in question. This

may be quite a different point in his development than the level at which the skill is first introduced, and the touchstone to tell us when this mastery stage should be reached may indeed be the point at which most pupils in some large unselected population answer a test question correctly. Perhaps a sensible percentage to denote mastery would be 75% right responses in a group having had equal exposure to instruction.

2. Along this same line, more and more communities will reject the idea of comparing their performance in terms of total score on a test which contains many items which they feel are undesirable because they do not want to include the content in their curriculum. This being the case, we will find, more, and more, that communities will ask for and get a comparison of their performance with national norms based upon the sub-set of items which they accept as measuring valid and pertinent goals of instruction within their own situation. The techniques for doing this exist at the present time and need only to be implemented. Once this fact is generally known, I think the demand for this kind of service will become much greater than it is at the present time.

3. Much greater attention will be paid to community factors which condition learning. In our present national standardization programs we are taking this into account in the way the strata are determined within which the participating school classes are randomized. We are now able to show on a bivariate-type chart the relationship between average community ability as measured by some school learning potential measure and the average achievement of school children in that community within each of the basic subject matter areas of reading, mathematics, etc. The existence and commonplace use of computers has made this type of comparison simple and the time has come when we have no excuse for not making such analyses routinely.

Thus, we will undoubtedly find many communities where the average level of learning potential is high, but the level of achievement, in some specific subject, is comparatively low, simply due to the fact that there is a wrong concept of the whole idea of being "at the norm." For example, it simply is not good enough for a community that is at the 85th percentile in terms of its average learning potential to be at the 50th percentile in terms of its measures of achievement.

4. Greater emphasis on the evaluation of community factors as conditioners of learning will bring about more and more attention to the need to equalize educational opportunities especially in the poor and deprived communities. This move will be effective in improving education in these communities only to the extent that these efforts receive broad community support. An anticipated difficulty that will almost surely condition gains in such communities is the low level of value put on an education by some groups. The notion of guaranteeing an "education" suited to the needs of the individual will surely prevail in the long run.

5. Under the circumstances as I have outlined them above, it will inevitably be true that there will be more and more emphasis on systematic longitudinal studies including the study of the growth potential compared to actual learning of individual students. Some such cumulative indices of school learning potential as I have described should become routine. Guidance, in the educational sense, must take such data into account, not by authoritatively prescribing the "next step" up the educational ladder, but by helping the individual and his family make use of such information in the decision-making process. Work along these lines is now entirely feasible with our computer facilities, but first some consistent type of pupil code to facilitate data-banking is an absolute necessity. There still will remain the tremendous task, once the data are systematically analyzed, of communicating this information to teachers and to parents as well as to students themselves.

Eventually all evaluative information will be open to the student and his parents without exception and a standard and essential part of the total school program will be the task of communicating with parents about all evaluative data, not just *occasionally* in terms of a rank card sent home every six weeks or so or a test profile. Frequent communication with the parent, sometimes in depth, but more usually to "update" the parent as to some difficulty the child may be having at the moment, will become routine.

6. Local test construction will become much more common. The task of writing good evaluative test questions is a difficult one, but this can easily be overcome by gradually accumulating for the use of the teacher a body of questions which can be drawn upon at his option. Portland, Oregon, for example is systematically attempting to do this, storing the test questions in a computer

in such a way that they may be called up on demand and put in form suitable for class use. This, in fact, is at the heart of criterion test development and interpretation.

7. Education in the near future will take much more advantage of audio-visual equipment and also computer aided instruction to provide for the presentation of learning materials to students with reading disabilities by means of an oral record. This can be done relatively simply such as by the use of cassettes in conjunction with 2" x 2" slides, or it can be much more complicated. We haven't begun to tackle the advantages of the cassette-slide combination in dealing with the slow learner, or in making possible more rapid progress by individuals who need to have access to instructional materials outside the range of the group to which the individual is normally assigned. This was the hope of programmed instruction in the first place and the failure of programmed instruction to make more of an impact on education, in my opinion, has been due to the fact that the equipment aspect was not up to the quality of the instructional materials.

Programmed instruction assumes a particular hierarchy through which the individual must go step by step whether he wants to or not. Branching is only a modest departure from this systematic presentation of materials. On the other hand, the use of cassettes and slides can be quite flexible since units concerned with a single topic provide the basis for the slides plus the cassette and these may be combined in endless ways.

Obviously there will be need for evaluating the amount of learning taking place under these circumstances. Locally made tests will be needed but eventually the standardized test which is a kind of generalized measure of the learning that has taken place as a result of *all* of the educational techniques employed will remain the bulwark of the evaluative program.

8. Non-test techniques of evaluation will gain in frequency of use and in efficiency. Presently these techniques are very hard to apply and efforts to develop rating scales, observational techniques, etc., have bogged down in the sheer complexity of getting teachers to pay attention to them in the midst of the total problem of instructing children. Such techniques must be devised in such a way that the student can evaluate himself without much attention on the part of the teacher.

In many of the junior colleges and community colleges, where there is a wide range both of ability and of interest in subject matter areas, the contract approach, not dissimilar in basic structure to the old Winnetka Plan, has been applied with success.

The greatest development of the elementary guidance program will be of assistance at this point. There may never be a satisfactory substitute for face-to-face conversation between an adult and a child who is having educational problems. It certainly still is the best way to diagnose and assess the needs of the child with regard to a changed learning situation.

9. Survey techniques like the National Assessment Program will develop and be used frequently where the sole purpose is to get an idea as to the level of learning reached by some representative groups, both in terms of mastery of specific questions, which is emphasized in the National Assessment Test, and in terms of comparisons with national norms. Survey tests which serve this only purpose can be much shorter than tests, like the present standardized tests, which serve both as survey measures as well as measures of individual pupil progress.

10. In order to implement any carefully worked out and well-integrated testing program covering a range of grades, it becomes absolutely essential to establish a pupil coding system which will permit the comparison of a child's performance year after year regardless of whether he changes schools within the system, or even if he goes from one system to another. The criterion here is that the coding system must be one that is constant for some substantially large population area which may be considered more or less homogeneous with respect to educational philosophy and curriculum.

It has been found in our studies in New Hampshire that such a code can be derived from the child's name, birthdate, and sex, but this is a rather cumbersome and difficult way of doing this job. Because the reason for repeatedly asking for birthdate information is not fully understood, there is some resentment over what appears to be needless duplication. Eventually it may be that Social Security numbers or something similar, will be assigned at birth and will become part of the nomenclature for a child within the school system just as much as his name is at the present time. When this becomes true, student records

will generally take on a much more realistic aspect, and it will become sensible, easy, convenient and effective, to study the child, not in terms of a single test representing a kind of snapshot of his performance but in terms of his profile of development over a period of years. Until we do reach this day, we can say that testing and evaluation in all its aspects is still in its infancy. The need for computer assistance along these lines for the storage and quick retrieval of data is obvious. As a matter of fact, more and more emphasis should be placed upon the use of computers for this purpose than on an increased use of these devices as aids to instruction.

11. In the 70's we will find increasing emphasis on the problems of the children comprising the middle 50% of the group. In the past, we have paid attention to the very slow learners such as the educable and trainable children, but we have done this because we have readily recognized them as outside of the bounds of normal status in society. We have also provided special help for the stanine 1, 2, 3, children through special classes, etc. I am concerned as well with the group above these levels but falling below the level at which the traditional college program makes any sense. The present almost universal goal of Americans seems to be to send their children to college. This is about as senseless as the former system which said that college education was the sole prerogative of those who happened to be born within the proper strata of society and at the top of the ability continuum. Post-secondary education for everyone who needs it to be trained for a job and trained to develop proper attitudes toward society will become commonplace. Community colleges, junior colleges, and the like will continue to spread rapidly, extending public education at least up through grades 13 and 14 with wide ranges of content in the curriculum. However, the high degree of specialization of learning that is represented by the liberal arts colleges and the associated colleges of engineering, business, and the like will always call for a facile mind which is not something to be developed without the potential for such development. Who is to say to what extent this potential resides in our present population? To say that we knew the answer to this would be to say that we have exhausted the opportunities to improve the educational process, an hypothesis with which I hope few would agree.

Yet it seems so patently evident that there are exceptional individuals who are able to learn, to create, to organize and disseminate knowledge of such complexity that it is far beyond the reach of the average person that this point of view is unassailable. For these gifted individuals also our educational system must be radically over-hauled and made much more effective than it is at the present time. Here again there is great need for more effective instruments both for identification and for evaluation. Better measures of creativity are needed. One might call them aptitude measures on a higher plane than those of the past. What indeed are the special qualifications and skills needed for outstanding success in medicine, science, both social and physical, art, literature, business, and government? What changes in curriculum (not in courses per se) are needed to foster development of these exceptional individuals, and what measures can be used to evaluate our success in their education?