DOCUMENT RESUME

ED 104 921 TM 004 377

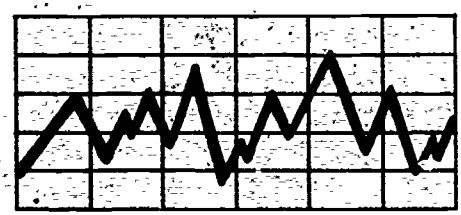
AUTHOR Womer, Frank B.
TITLE Test Norms: Their Use and Interpretation.
PUB DATE [65]
NOTE 62p.
AVAILABLE FROM National Association of Secondary School Principals, 1201 16th Street, N.W., Washington, D.C. 20036 ($2.00 per copy; discount: 2-9 copies, 10%; 10 or more, 20%)

EDRS PRICE MF-$0.76 HC Not Available from EDRS..PLUS POSTAGE
DESCRIPTORS Comparative Testing; Group Norms; *National Norms; *Norm Referenced Tests; *Norms; *Standardized Tests; Testing; Testing Problems; *Test Interpretation

ABSTRACT

The process of determining test norms and of using them realistically is discussed. This process involves putting meaning into a single test score by relating it to other test scores achieved by other pupils or by the same pupil at other times. The focus of attention is on the meaning of test norms; it is not on the meaning of types of test scores. For example, how norm tables that yield IQ's are developed will be considered, but not the concept of the IQ itself. Some attention is given to the actual development of percentile norms, but the various characteristics and facets of percentiles are not discussed. Attention is centered around norms for widely used, nationally standardized tests, as distinct from teacher-made classroom tests. Some of the principles discussed apply to both, but the major purpose of this publication is clarification of national test norms. (Author/BJG)

TEST NORMS
THEIR USE AND INTERPRETATION



Frank B. Womer

National Association of Secondary-School Principals

THE NATIONAL ASSOCIATION OF SECONDARY-SCHOOL PRINCIPALS
1201 Sixteenth Street, N.W.
Washington, D.C. 20036

Two dollars per copy
Quantity discounts: 2-9 copies, 10%; 10 or more copies, 20%

3/4

Contents

Foreword

What is a norm? How does it affect the preparation and scoring of standardized tests?

Adequate replies to these questions imply a sophistication that has yet to be achieved by students, parents, and teachers. The experts agree that the mystery of "norming" a test can be dispelled only by a clearer understanding of how norms affect the interpretation of test scores. That is why the National Association of Secondary-School Principals was advised to prepare a plainly written analysis of the importance of norms.

In the search for someone who could clarify concepts and procedures about norms and norming techniques, we asked Frank Womer of the University of Michigan to tackle the job. With a minimum of technical fuss, he shows how norms are arrived at and how they should be reviewed. We hope that his work will lead to increased clarity not only within the profession but also among the intelligent, devoted citizens whose support is the cornerstone of all educational progress.

ELLSWORTH TOMPKINS
Executive Secretary
National Association of
Secondary-School Principals

What This Booklet Is All About

"What a big boy he is!" is a well worn phrase—used so often one might almost assume it had a common meaning for everybody. Yet we all know its real meaning depends largely on when or where it is used. If someone says it as he looks down into a baby carriage, it probably means the boy is physically big in relation to other babies his own age. If it is said by a visitor who has not seen the boy for a while, it may only mean that he is surprisingly bigger than he was a year ago. If it is said by the football coach as he looks over prospective tackles on the college freshman team, it may mean that the boy is already bigger than the typical college tackle. In each context the boy is big only in comparison with some standard of bigness—some norm—whether that norm be babies in general, his own former size, or the size of football tackles.

In these pages we shall be talking a great deal about test scores and norms. And the very first point we wish to make is that they, too, take their meaning largely from some kind of comparison. A pupil's test score is high or low only in relation to other pupils' scores, to his own previous scores, to a selection score indicating possible success in a certain college, or to some other criterion or standard for judging what is high or low.

This discussion deals with the process of determining test norms and of using them realistically. It deals, then, with the process of putting meaning into a single test score by relating it to other test scores achieved by other pupils or by the same pupil at other times. The focus of attention is on the meaning of test *norms;* it is not on the meaning of *types* of test scores. For example, how norm tables that yield IQ's are developed will be considered, but not the concept of the IQ itself. Some attention will be given to the actual development of percentile norms, but the various characteristics and facets of percentiles will not be discussed.

Another delimiting feature of this brochure is that the discussion is centered around norms for widely used, nationally standardized tests, as distinct from teacher-made classroom tests. Some of the principles discussed apply to both, but the major purpose of this publication is clarification of national test norms.

Some Preliminary Thoughts on the Purposes of Norms

To have an example by which to illustrate various aspects of test norms, let us run through the process of developing a standardized test. Call it the "XYZ Arithmetic Test," for grades 7, 8, and 9. Assume that it consists of 25 multiple-choice items, that these were written to sample the arithmetic knowledges and skills one can reasonably expect junior high pupils to have developed, and that a good job of item writing was done.

Let's say that this test was given to John Smith, an 8th grader, and that John answered 16 of the items correctly. From this information alone all we can say is that John got 16 items correct out of 25 possible. If we want to know anything about how John's performance compares with that of other eighth graders, we need information about the scores of other eighth graders. If we want to know how John's performance compared with his performance on a similar arithmetic test he took in grade 6, we need information as to his relative performance on that sixth grade test. If we want to know whether John's arithmetic test score is high enough to suggest that he might be successful in algebra next year, we need to know the scores of other pupils who later elected algebra and succeeded in it.

To make significant judgments about John's score on the XYZ Arithmetic Test we must know the scores of other pupils on the same test. We need yardsticks based on the performance of many pupils so that we can place John's score on those yardsticks. In our example, we need three yardsticks: 1) the performance of other eighth graders on the XYZ Arithmetic Test; 2) the performance of other sixth graders on the same arithmetic test that John took in grade 6; and 3) the performance on the XYZ Test of eighth graders who subsequently elected algebra and were successful in it. The entire process of developing test norms is designed to provide yardsticks, so that the test scores of an individual or group can be compared to the performance of comparable individuals or groups.

Test norms provide check points of performance in relation to other pupils—they say that John scored higher (or lower) than other eighth graders, that he scored higher (or lower) than other pupils succeeding in algebra the following year. Test norms do not tell whether John's performance was as good as his teacher might have desired, or whether he had mastered every concept in arithmetic to which eighth graders are exposed. They simply tell us where John stands in relation to the performance of "known" groups of pupils.

Similarly, a measurement on a yardstick does not, in itself, establish tallness or shortness. It merely establishes height. A ten-year-old boy whose height is five feet is considered tall, but an adult man whose height is five feet is considered short. In like fashion, a score of 16 on the XYZ Arithmetic Test may be average for eighth graders but above average for seventh graders. The purpose of test norms is to add meaning to individual test scores in the same way that meaning is added to height by judging tallness in relation to some known group of heights.

A Working Example of the Development of Norms

We said we would assume that a good job of item writing was done for the XYZ Arithmetic Test. We need also to assume that more than 25 items were written originally and that the larger group of arithmetic items had undergone a process called item analysis. This is done by "trying" the items on a representative group of students. Item analysis helps to determine which items in a test are good ones. "Good" items are those which are answered correctly by most pupils receiving high scores on the total set of items, answered correctly by about half of the pupils receiving average scores, and answered correctly by only a few pupils receiving low scores. Thus, on an overall basis, good items are missed by about as many pupils as those that get them correct. Good items tend to separate pupils in the same way that the total test separates them; that, is they *discriminate* well between the better students and the poorer ones. Poor items are those that are answered correctly by almost as many pupils receiving low scores as by those receiving high scores on the total test or set of items. Sometime a very poor item is answered successfully more often by poor students than by good ones; it discriminates negatively. If one wants to be sure of having 25 good items after item analysis, it is necessary to write and try

out perhaps as many as 50 items. Once a group of items is chosen, they are put together as a test, directions for administration are developed, and the test is ready to be standardized or normed.

Arranging the Sample

The first step in the standardization process is to define the group (the "population") of schools or pupils for which the test is presumably appropriate. If the XYZ Arithmetic Test was developed so that it might be useful in any junior high school in the United States, then we can define its population as all junior high pupils (grades 7, 8, 9) in the United States. If the test had been designed primarily for use in independent, junior high schools, we would define its population as all junior high pupils enrolled in independent schools. If it had been designed for a state testing program in Oklahoma, the population would be all junior high pupils enrolled in Oklahoma schools. The definition of the appropriate population for a test is dependent upon the objectives of the author and the projected use of his test.

Once a test population is defined, the next step is to select a sample group of schools or pupils from the defined population. In theory it would be ideal to select a random sample of pupils for a nationally standardized test. In practice this is not done because of the difficulties of securing a sample consisting of a single pupil in some schools, two pupils in many others, three in still other schools, and so on. It would be very difficult to organize such a standardization even if one could afford the enormous amount of money that it would cost. In practice, then, school systems generally are used as samples, and attempts are made to see that each school system has an appropriate (not equal) probability of being chosen. For example, a junior high school with an enrollment of 1,000 should have a greater chance of being in the sample than one with an enrollment of

100, so that each of the 1,000 pupils would have the same chance as each of the 100.

Schools generally are classified in two ways before the random selection is made. First, since there is considerable evidence that average pupil performance is a bit higher in some geographic areas than in others, most samples of school systems for test norming are chosen so that all geographic areas are represented. Sometimes each state is represented. The second major classification is according to size of school system (or community). Again, average test performance varies somewhat from size to size of community. It is felt that large communities, middle-sized communities, and small communities all should be sampled in a norm group.

Putting all this together, we find that the author of our XYZ Arithmetic Test proceeded as follows:

1. He defined his population as all junior high pupils (grades 7, 8, 9) in the United States.
2. From census data he listed all school systems separately by state.
3. Within each of the 50 state groups he separated communities into four categories by population: (a) over 500,000; (b) 50,000 to 500,000; (c) 5,000 to 50,000; and (d) below 5,000. This produced a few less than 200 groups (some states have no communities over 500,000).
4. Within each of the nearly 200 groups, he chose, at random, one or more systems, depending upon the total population in each group. Let's assume that he selected 300 systems.

Remember that this example is for illustrative purposes and should not be interpreted too literally. In practice it might not be necessary to sample every state; one might choose different geographical divisions. In practice one might also treat certain communities in a special way (those over 1,000,000 for example).

One could also cite examples of test standardization geared to socio-economic level of community, or examples in which specific school buildings (not systems) were chosen. But, regardless of details, the sample chosen for the XYZ Test would be fairly typical of samples chosen for tests designed to be used nationally. Each sampling is an attempt to come as close as possible to giving each pupil in the defined population an equal (random) chance to be selected.

Working with the Schools

Once a norming sample has been selected, the next step is to seek the cooperation of the chosen schools. In our example we theoretically ended up with 300 school systems. The next thing to be done would be to write a letter to the superintendent of schools of each system (possibly with a carbon copy to each building principal having grades 7 and/or 8 and/or 9). The letter would describe the XYZ Arithmetic Test, would point up the need for a new arithmetic test and the careful construction of the XYZ Test. Further, it might point up the importance of local school systems in the continuing research on new tests and the distinction of being one of only 300 school systems chosen. The letter would certainly offer to report all pupil scores back to the school. It might or might not offer the tests themselves to the participating school, and it might or might not suggest a modest fee for participation in the norming program. In any event, it would ask the superintendent and/or principal(s) to agree to test their pupils with the XYZ Arithmetic Test.

Now let us assume that in response to the initial appeal, 180 (60 per cent) agree to participate by testing their pupils with the XYZ Test. The question of adequacy of a 60 per cent sample is the next point to consider. If one can assume that the 40 per cent who say "No," or who don't even acknowledge the request, are no different from the 60 per cent who say "Yes," one can proceed to the actual testing. But, if one makes the assumption that the 60 per cent tend to be systems that are a bit larger and

more dedicated to research, represent a somewhat higher socio-economic level, with slightly abler students than the 40 per cent, one is apt to be closer to the truth.

At this point several choices are open. One can proceed directly to testing, hoping that the 60 per cent will not be too atypical of all schools. One can try to convince the "No" schools to say "Yes." Or one can attempt to secure the cooperation of other school systems like the ones that responded negatively, by choosing, at random, a substitute for each. One way or another, attempts are generally made to move nearer to a complete sampling. How far one carries these efforts is determined by the test author and his publishers. Let's assume that, in the case of the XYZ Test, substitute schools were chosen and half of the substitutes agreed to cooperate. This would mean that 240 (180 plus 60) school systems actually agreed to test their pupils—80 per cent of the sought-for 300.

No matter how far a test author pushes his sampling procedures, he is not likely to secure 100 per cent participation. Perhaps the best example of participation that one can point to is Project Talent. In that widely publicized research study approximately 95 per cent of the school systems selected did do the subsequent testing. Since most nationally standardized tests do not achieve so high a proportion, one can be sure that some error of measurement is present in each test standardization because of imperfect sampling. (Some error of measurement is inevitable because of the use of a sample rather than the entire population.) Some of the variation among norms of nationally standardized tests can be traced to the fact that no one does a perfect job of sampling a national population. Thus, it is asking too much of different tests to expect the resulting scores to be exactly interchangeable.

The next step is the actual administration of the test. In this phase many teachers and counselors are apt to be involved. In most cases the test administration is quite like that done in a school's on-going testing program. The test publisher establishes

approximate dates for test administration; the local school picks an exact date. The test publisher sends sufficient test booklets, answer sheets, special pencils (if necessary), and directions for administration to each school; the school administers the test according to prescribed directions and returns all answer sheets to the publisher. The publisher eventually reports pupil scores back to each school. (This last step is irrelevant to the test standardization itself. It is done for public relations purposes and to provide each school an incentive to participate.)

At this point one assumes that the test has been administered uniformly in the various schools. One assumes that each teacher or counselor or administrator who gave the test did follow the printed instructions to the letter. This probably is a reasonable assumption for most schools. Fortunately, if and when errors do occur in the test administration, they will not necessarily affect the norms, *if* part of the errors are of the type that would raise test scores (such as not calling time exactly and letting a group have an extra minute or two on a test) while others are of the type that would lower test scores (such as failing to read directions accurately so that some pupils may not know exactly how to answer the questions). Errors of administration would be a major factor only if they were all, or almost all, operating in the same direction (either to raise scores generally or to lower them generally).

In any event it is undoubtedly true that some variability in test scores can be traced to test administration. This, again, is a factor that may cause some variation in norms between two tests purporting to measure the same thing.

Organizing the Scores for Interpretation

After our XYZ Arithmetic Test has been administered in the 240 school systems, let's assume that 20,000 pupils were tested at each grade level, 7, 8, and 9. Now what happens to those 60,000 test scores? Or, more specifically, what happens to the 20,000 scores for grade 8 pupils? Let's assume that one wants

to use percentile ranks for reporting the results of the XYZ Test. In all likelihood an electronic computer would be used to compute the percentile ranks, and provide a set of percentile ranks such as the one given in Table 1.

Table 1
XYZ Arithmetic Test
Norms for Grade 8

Raw Score*	Percentile
23-25	99
21-22	95
19-20	90
18	75
17	70
16	60
15	50
14	40
13	30
11-12	25
8-10	10
6-7	5
0-5	1

* Since most pupils get middle scores and few get very high or low scores, several different raw scores sometimes yield the same percentile.

This make-believe table illustrates one simple way of reporting test norms. Examples of interpretation would be:

1. Mary got 18 items correct, giving her a percentile rank of 75. This means that, of the 20,000 eighth graders who took this test during the standardization process, three out of four got scores lower than Mary's.
2. Bill got 13 items correct, giving him a percentile rank of 30. This means that, of the 20,000 eighth graders who took this test during the standardization process, only 3 out of 10 got raw scores lower than Bill's.

If one is willing to assume that the pupils tested in the XYZ Arithmetic Test standardization are typical of all eighth graders nationally, one can extend these sample interpretations to read:

1. Mary scored higher than 75 pupils in a group of 100 typical eighth graders.
2. Bill scored higher than 30 pupils in a group of 100 typical eighth graders.

Norms and Standards

We shall return to the question of interpreting test norms in a later section, but now let's look again at the XYZ norms. Think of the XYZ Test as one using multiple-choice type items with five choices (a, b, c, d, e). A pupil not knowing the answer to a particular question could guess at the answer, and every so often (about one time in five) he would be likely to get an item correct by chance alone. In a test of 25 items of that type a person knowing absolutely nothing about eighth grade arithmetic might get a raw score of 5 (or 4 or 6) by chance alone. Thus, in a very real sense, the lowest "theoretical" score is 5 rather than zero. Or one might choose to say that the *effective* range of scores on the XYZ Test is from 5 to 25.*

All of this is leading up to another point, a point that differentiates standardized test norms as yardsticks from the yardsticks that teachers generally use with their own classroom tests. Notice that a 50th percentile is achieved by any pupil getting a raw score of 15 on the XYZ Test. But 15 is only halfway between the lowest theoretical score and the highest (5 to 25). This says that an "average" pupil on the XYZ Arithmetic Test actually got only half of the items correct above the chance level. What self-respecting classroom teacher would classify as average a pupil getting only half of the items correct on one of his tests?

Most classroom teachers giving classroom tests would expect their "average" pupils to get at least 80 or 85 per cent of the items correct, because "failure" (as defined in many schools) is represented by anything less than 70 per cent correct response.

* In practice some pupils do get less than chance scores. This may be due to misinformation—to a pupil's having learned wrong concepts rather than correct ones. The chance scores apply only when the pupil makes pure guesses.

Such a yardstick, if applied to the XYZ Test, would suggest that "passing" (minimum accepted raw score) would be 18, and "average" would be a raw score of 21. But the typical teacher's yardstick of pass or fail as associated with certain percentages is *not* the yardstick of test norms. *The yardstick of test norms is based entirely upon actual performance of pupils and not upon any predetermined level or levels of performance.*

A Look at Some General Characteristics of Test Norms

When items were selected for the XYZ Arithmetic Test, they were not selected as a teacher selects (or writes) his. The use of item analysis procedures (see page 4) produces items that, as a whole, are more difficult than a pupil would face in a classroom test. In fact, the items are chosen carefully so that the average score, the one corresponding to the 50th percentile, will be rather close to the middle of the range of possible scores. The author does this so that the scores on his test will be spread out from highest possible to lowest possible. For example, suppose that the XYZ Test had produced scores ranging only from 15 to 25 rather than from 4 to 25. It would then provide only eleven raw score possibilities, instead of twenty-two. Spreading the scores out allows for finer distinctions between performance of pupils, which produces greater accuracy, which is what we mean by test reliability.* A classroom teacher is most interested in discovering which of his pupils have mastered a given assignment or developed certain skills at an acceptable level. The author of

* Reliability may be defined as the extent to which the scores on a test are free from being influenced by errors of measurement. In our example the reliability of the XYZ Test would be the extent to which pupils' true knowledge of arithmetic is reflected in their test scores, as against the extent to which any errors of administration or scoring or variatons in motivation, etc., entered into their scores.

a standardized test is most interested in putting pupils in rank order on the skill being measured.

Norms Rank Order Pupils

This simple fact of test norms is commonplace to test technicians. Yet it is easy to miss or forget the distinct difference between judging pupils against a yardstick of mastery of content and judging them against the yardstick of performance of peer groups of pupils.

This concept applies to ability tests as well as to achievement tests. An IQ of 100 says simply that a pupil has performed on a general ability test at a level higher than half of a group of typical pupils his own age. There is no absolute scale (yardstick) of intelligence against which a pupil can be measured.

The example of norms for the XYZ Arithmetic Test was developed using percentile ranks. It could have been developed using T-scores or stanines or band scores or even grade equivalent scores.* All are derived scores, and there is no one type of derived score that *must* be used for a particular test. Statistically, one could even develop an IQ scale for our XYZ Arithmetic Test. Hopefully, no one with common sense would do it. The choice of type of norm to be used to report the scores for a given standardized test is up to the author and publisher. In practice the authors of achievement tests designed for the elementary grades generally choose to develop norms in terms of grade equivalent scores and percentiles. Authors of intelligence tests generally choose to use an IQ scale (though some use per-

* T scores: A type of derived score ranging from 20 to 80, with an average score of 50. It generally presupposes a normal distribution of scores.

Stanine: A type of derived score ranging from 1 to 9, with an average score of 5. It assumes a normal distribution of scores.

Band scores: A type of derived score giving a range rather than a single point, e.g., a percentile band of 40-60 rather than a percentile of 50.

Grade equivalent score: A type of derived score based upon the average performance of pupils enrolled in specific grades; e.g., a grade equivalent of 8.5 means test performance comparable to that of a pupil in the fifth month of the eighth grade.

centiles or percentile bands). Authors of aptitude tests and other secondary level tests often choose percentiles, but some now are reporting test scores as stanines. An organization such as the College Entrance Examination Board may choose to develop its own score scale (from 200 to 800, with an average score of 500). Whatever the choice is, it probably will be based on precedents and author preference. The choice is not apt to be dictated by statistical or normative considerations.

Norms May Be National or More Limited

The question of whether to report test norms as *national* norms also is optional with a test author. Most tests used in the schools do have, or claim to have, national norms. This is based on custom and on a desire of the author and publisher to sell their product in all parts of the country. A test with norms based only on pupils from Maine is not apt to sell widely in California. This is not because the norms would necessarily be inappropriate, but there certainly would be an element of doubt in the minds of the potential users. Test users tend to feel more confidence in a test if they know that students in their own state or region were included in the norm group.

The user of any standardized test will almost certainly be provided with national norms. Occasionally he also may be able to obtain regional norms from the test publisher. In some cases he can get state norms from a state testing program operating in his own state. If he is ambitious and has a bent toward working with numbers, he may develop his own local norms or cooperate with other schools in developing group norms. Whatever norm group he chooses to use, individual pupil results must be interpreted in relation to the population of schools in that norm group. While the user of national norms may rightfully compare a pupil's performance to that of other pupils his age or grade nationally, the user of state or local norms must make his interpretation by comparison to state or local groups.

Test Norms Are Not Absolute

By now the reader should be well aware of the relative nature of test norms, which grows out of relating each pupil's score to the scores of other pupils who took the same test. There are also other factors which tend to make test scores relative rather than absolute.

For example, our hypothetical XYZ Test was designed to sample arithmetic skills commonly taught in junior high schools in this country. But not all junior high schools expose pupils to the same arithmetic skills at the same level. Many schools are teaching some form of 'modern mathematics." To the extent that a particular school system may be covering some aspect of modern math *over and above* the traditional arithmetic skills, its pupils should be able to perform satisfactorily on the XYZ Test. Even in these cases, however, the pupils will not be able to demonstrate their added proficiency in modern math on the XYZ Test. Furthermore, some school systems with modern math programs may not cover the traditional arithmetic skills to the same degree that other schools do; thus, their pupils may not perform as well on the XYZ Test as pupils from these other schools even though they are equally good at arithmetic as a whole. Applying the same norms to students in different schools is justified only if they have had essentially the same opportunities to learn what is being tested.

The relation of test norms to content applies also to standardized tests of ability. No one has ever established an absolute scale or test to measure general intelligence—a test that is not at all dependent upon what the pupil taking the test has learned. If different children of the same age have not had similar opportunities to learn, then differences in their scores may be due to differences in opportunity rather than in ability. Furthermore, authors of different intelligence tests sample somewhat different aspects of intelligence. One test author may choose to measure verbal ability by using verbal analogy items:

(*Cold* is to *hot* as *straight* is to ___. a. crooked b. narrow c. up d. out e. forward;

another by using regular vocabularly items; and still another by sentence completion items:

(What goes up must come ___. a. down b. flying c. crashing d. a cropper).

While each type of items may very well measure some aspect of verbal ability, the various types do not necessarily measure identical elements of that ability.

Test Norms Are not Universal

Perhaps the most obvious way to illustrate this lack of universality is to ask: Would it be feasible to administer our XYZ Arithmetic Test to pupils in France? On straight computation items we could expect French pupils to perform satisfactorily, but on story problems printed in English not at all well. The norm group f r the XYZ Test was based on junior high pupils in the United States and it is unrealistic to assume that pupils everywhere have been exposed to exactly the same skills and understandings and will perform exactly the same way.

Even in the domain of ability testing, where one might choose to use some pictorial type of item:

(△→△ : : □→__a. □ b. □ c. △ d. ○ e. □),

it is unrealistic to assume that pupils everywhere have learned geometric figures and relationships between them in exactly the same way and at exactly the same ages.

Test Norms Are Not Permanent

Test norms are a product of the time when they are developed. If a test is used with the same pupils at two widely separated times, we expect their scores to change. Probably the most obvious example of changes in norms as pupils age is that provided by the fall and spring norms developed for most elementary-level achievement tests. On any achievement test, the

norms for the end of a particular grade may be expected to be higher than for the beginning of that grade.

Thus, any standardized test that is to be used at more than one time during the age or grade development of pupils must have separate norms for each time, or time unit, when it is to be used. A test such as an algebra aptitude test may get by with only one set of norms—applicable at the point just prior to beginning formal instruction in algebra, say the spring term of the 8th grade in most schools. But our XYZ Arithmetic Test, designed for grades 7, 8, and 9 would need a minimum of three sets of norms and preferably six (both first and second semester norms for each grade).

There is another important factor which makes norms a product of a given time period. Human knowledges and understandings themselves change over time. For example, a pictorial intelligence test developed in 1930 might have used a picture of a trolley car, a refrigerator with the motor on top, or an airplane with two wings, one above the other. All of these pictures would have been recognized easily by boys and girls in the 1930's, but could be puzzling for some children today. In like fashion, our language itself changes. At one time the word Mars had its traditional meanings when used as a vocabulary item in a test. Now it must also be keyed correct if the answer is candy bar.

Furthermore, school programs and teaching methods change. Particular subject matter may be introduced earlier, the school may become more demanding, improved instructional materials may make learning easier. Such changes can have considerable effects upon pupil scores over a period of years.

Because of these changes in knowledges and understandings and ways of organizing instruction, tests must be re-examined periodically, say every five years or so. A thorough re-examination requires two things: 1) a statistical check of each item in a test to see whether pupils still answer it in the same way they did previously; and 2) the development of new norms for the entire test (or a revised version) so that any overall changes in

group scores are reflected in the norms. Item analyses generally are not undertaken in less than an eight or ten year span and sometimes not that often. The development of new norms may take place more often, particularly if evidence begins to accumulate that a set of norms is for some reason unrealistic.

In any event, it is important for a test user to ascertain the time when the norms were established for the test he is using and to interpret results accordingly. If the norms of our XYZ Test were developed in 1964, and those of another similar arithmetic test in 1954, a user might question whether results from the two tests should be compared. The many changes taking place in mathematics instruction in the intervening period could have affected the two sets of norms.

Another very important consequence of the fact that test norms are a product of a particular time relates to the question of comparing test results taken from different tests at rather widely separated points in time. The apparently simple question of whether pupils could read better twenty years ago, or compute better, or do anything else better is almost impossible to answer exactly by using test results. Over a twenty year span of time a particular test might well have been revised and supplied with new norms twice or more. Thus, two different yardsticks would be used to assess similarities or differences over such a long period of time.

One can develop an example of this phenomenon with our XYZ Arithmetic Test. We hypothesized a set of 1964 norms with a percentile of 50 corresponding to a raw score of 15 for grade 8. Suppose that in 1974 new norms were to be developed and at that time a raw score of 17 turned out to be average. We would then assign a percentile rank of 50 to the raw score of 17 and change all of the other percentiles appropriately. Thus, for exactly the same set of test items, a pupil in 1974 would have to get two more items correct than a pupil in 1964 in order to achieve the same relative position (50th percentile) among 8th grade pupils.

It cannot be stressed too often that test norms are yardsticks developed at a particular time, with a particular group of pupils, and with a particular selection of test content. To generalize to other times, to other types of pupils, or to other content is a questionable practice.

Test Norms Assume Comparable Educational Background for All

In a strict sense this statement is not true, for it is test authors and test users who make this assumption rather than the norms. If for our XYZ Test norms we had used only a thousand 8th grade pupils, all of whom attended one selective junior high school, the mechanical procedure of developing the norms would not have changed. However, our purpose in doing such a thing would have been different from our purpose in sampling all 8th graders nationally, and users' interpretations would be vastly different. In general, a test designed to be given nationally includes items that sample understandings to which all or almost all pupils have been exposed. The reason some pupils score higher and others lower is, then, primarily dependent upon each individual's retention of knowledge or level of development. The difference is assumed not to be the result of differences in opportunity to learn.

In like fashion, intelligence tests assume a common background of opportunities to pick up information and skills, some related rather directly to school activities, others related to learnings acquired outside of school. The assumption is made that the more intelligent pupil will develop his knowledge and skills to a higher level than the less intelligent pupil, both having been exposed to comparable educational opportunities in and out of school.

Clearly, then, an atypical pupil or an atypical group of pupils (atypical in terms of educational opportunities) may not be judged "fairly" by a test which assumes equal educational backgrounds. Some may have had very rich opportunities to learn,

others very meager ones. The differences among their scores—or some part of the differences—may, then, be chargeable to differences in opportunity rather than in ability.

However, before taking this statement as a condemnation of tests or of test forms, remember what the basic purpose of test norms is. By definition, norms are yardsticks designed to relate individual pupil performance to the performance of "known" groups of pupils. The known groups generally are and should be "typical" groups. If a test is designed for widespread national use it must be geared to the average or typical population.

To reconcile the fact that test norms are geared to the general, typical pupil population while some pupils or groups of pupils are not typical calls for common sense on the part of test users. To condemn test norms for not providing useful information for all pupils under all conditions is unrealistic, just as it is unrealistic to condemn a textbook in reading designed for 5th grade pupils because there are some 5th graders who cannot grasp the content while some others are far ahead of the text.

How might all of this apply to the XYZ Arithmetic Test? We hypothesized that its norms were designed for use with all 7th, 8th, and 9th graders nationally. Thus, it was designed to measure arithmetic knowledges and understandings common to typical junior high pupils in the United States. But one can see how judiciously its results must be interpreted if we look at two examples:

1. John's father is an engineer who works constantly with figures. He has a desk calculator in his home and has taught John how to use it. In fact, John does some of his father's calculations for him and receives a small stipend for his work. John's teachers know of his work with figures and have encouraged and utilized his arithmetic skills. John scored at the 95th percentile on the XYZ Test while in the 8th grade.

The only fact shown by the test results is that John scored higher than 95 per cent of other 8th graders; the reasons why he

scored that way are speculative. Perhaps he is bright, perhaps he studies hard, perhaps his added practice at home increased his score, perhaps all of these factors are present.

2. Tom's father is an itinerant field worker, whose formal education stopped at grade 6. Sometimes his family is with him, sometimes it is not. Tom often starts school in September in one community, but when his family moves in October he is out of school for a few weeks and then enters a different one. Only the simplest arithmetic, mostly making change, is used in his home. Tom learns and practices arithmetic skills only in school. Tom scored at the 25th percentile on the XYZ Test in grade 8. What does this mean?

Perhaps he really has less than average ability to develop his arithmetic skills, perhaps he is not highly motivated to do school work, perhaps his family background is a handicap he has not been able to overcome. If Tom had been reared in a typical home with typical educational advantages he might have scored even higher on the XYZ Test than John did. Nevertheless, it is a fact that, when compared with an average group of 8th graders on this arithmetic test, his score was surpassed by 75 per cent of them. Again, common sense must be applied to determine the why or whys.

Test norms establish relative position within a group. They do not establish the reasons for that position. A pupil with an atypical educational or sociological background may score higher or lower on a test than he would have under typical conditions.

Test Norms Are More Apt to Reflect Typical Performance Than Maximum Performance

There is a notion about pupil performance on standardized tests that says, "You can't score higher than your actual ability (achievement) level." Such a statement seems so obvious that few would question it. Actually it should be questioned, for at least two reasons: 1) the choice-type of test item does allow a

pupil to receive a higher test score by chance alone than his theoretical "true" score; and 2) the process of developing test norms suggests that typical performance rather than maximum performance is sampled.

If one remembers that most standardized tests now use multiple-choice items, it follows that most pupils will guess on some or all of those test items for which they do not really know the answer. Most test authors no longer use a correction-for-guessing formula. They are more apt to suggest that all pupils attempt to answer all items. This is an invitation to guess when the answer is unknown. For most pupils guessing does not alter their relative performance appreciably. This is because, if everyone guesses, all raw scores are raised slightly, and the rank order of the scores remains essentially the same. (Remember that the major purpose of test norms is to determine rank order.) Occasionally, however, a pupil may suffer or gain by guessing very poorly or very successfully. Thus, occasionally, by chance alone, a pupil may get a lower score (relatively) than his real ability or knowledge warrants. And occasionally, by chance alone, a pupil may get a higher score (relatively) than his real ability or knowledge warrants.

The second point, relating to typical or maximum performance, probably is much more important than the first. There is a considerable body of evidence that demonstrates the importance of pupil motivation in test taking. Artificial means, such as rewarding testees with small cash payments, have been shown to raise group test results above those obtained under normal, routine testing conditions. Probably every reader can think of examples of pupils who "don't care" when taking tests and who seem to achieve lower test scores than their other behavior would suggest. And there is always the eager pupil who considers every test a personal challenge to get a high score. Thus, one can think of situations in which pupils may perform either better or worse than their own typical performance.

Think again of the standardization of the XYZ Test. It was given to 60,000 pupils in 240 different schools. It was handled by teachers, administrators, and counselors, undoubtedly with some variation in adequacy of administration. The pupils in this hypothetical case would have been told to do their best work, but that the test scores would not be used to grade them or in any other way be used for selection, or promotion; or honors, or other classifications. Some of the pupils in the 60,000 undoubtedly would take the testing very seriously and would work very hard at it; others would take a so-what attitude and just go through the motions, doing the easy items and ignoring or guessing at the difficult ones. But the large majority of the 60,000 probably would approach the XYZ Test with an average or typical amount of motivation—they would do their best, within limits. While this is speculative, it seems likely that the highly motivated and poorly motivated would tend to even each other out and the large number of pupils with average motivation would tend to dominate the scores in the norm group.

All of this rationale is designed to demonstrate once more the statement that test norms represent typical performance rather than maximum performance. Most pupils, under conditions of strong motivation can perform or achieve at a level above their day-to-day performance. Test norms are geared more to their day-to-day operating level of efficiency than to their maximum level.

Factors Affecting a Choice Between Specific, Well-Defined Norms and General Norms

Most of the examples used so far have involved national norms, although regional, state, and local norms have been mentioned. The only real difference between these various types of norms is the geographical definition of the population that is to be sampled. If there are differences between the national norms and the state norms for a given test, they are the result of achievement or ability differences of the pupils in that state when related to the other 49 states. However, there are times when it is desirable to develop and use norms based on or related to characteristics that are not dependent upon geography.

1. *Regional differences.* Before discussing other types of norms it might be well to mention a characteristic of national, regional, and state norms that has become clear in the past several years. There is considerable evidence to indicate rather consistent differences in group test performance between three large geographical areas of the country. Consider Area I as New England and the Middle Atlantic states (including Maryland on the south, and Pennsylvania and New York on the west).

Consider Area II as the Southeastern states (including all the so-called "border" states on the north and going as far west as the Mississippi River). Consider Area III as all other states, Midwest, Rocky Mountain, Southwest, and Far West.

Having defined these areas one can say that group test results on achievement and ability tests tend to be highest in Area I, lowest in Area II, and in the middle in Area III. These group differences are large enough to appear to be significant (non-chance). However, there is much more overlap than there is difference. Recent evidence for this statement comes from Project Talent[1] results and also from the selection scores for various states in the National Merit Scholarship competition.[2] Information from both sources shows the same regional differences. Project Talent results are from a wide-range series of general information tests, and data were gathered from a sample of all pupils enrolled in the 9th, 10th, 11th, and 12th grades in the entire country. National Merit selection scores are for a test of general educational development and apply only to the top one per cent of the pupil population. The reason or reasons for these differences are not easily determined; the reader must decide for himself why they appear.

Lest the reader be left with the impression that regional differences in norms are overwhelming, it should be mentioned that there are other differences between the achievement levels of various schools that are much more dramatic than regional differences. Project Talent results, for example, demonstrate that achievement in schools located in high socio-economic level areas is more like achievement in schools in similar socio-economic areas all over the country than it is like achievement in other schools in the same geographic area. The Project Talent staff developed a classification of schools according to the average

[1] Darley, John T. "A System for Classifying Public Schools" (Project Talent Results of Initial Analyses). A paper presented at AERA and AASA Meeting, Atlantic City, New Jersey, February 1962.

[2] *Guide to the National Merit Scholarship Corporation*, Evanston, Illinois, August 1963, pp. 13-14.

achievement level of their pupils. The classification is based on geography, socio-economic level, and size of community.

2. *Sex differences.* Many standardized tests (most general intelligence tests, achievement and skills batteries) make no differentiation in norms by sex. A few have developed separate norm tables for boys and girls. There is nothing particularly mysterious about sex differences in test norms. In general, girls tend to get somewhat higher scores on tests that are verbal in nature, that depend in large part upon command of the language. In general, boys tend to get somewhat higher scores on tests that are numerical or mechanical in nature. When separate norms are used for boys and girls, the person interpreting test results must make a mental note that he is comparing a pupil's score with scores of boys or girls only, as the case may be. When separate norms are not developed, there still may be some sex differences, but either they are small or the test author feels that it is best to compare all pupils on the same norms regardless of the difference.

3. *Types of school.* The independent schools (and certain public schools) associated with the Educational Records Bureau have felt a need for their own separate norms. These schools are selective in nature, tending to enroll pupils of above average academic ability. They have relatively homogeneous student bodies with similar abilities and similar goals. Since these student bodies are not typical of most public schools, many of the independent schools prefer to compare the performance of their pupils with that of pupils in other like schools. The Educational Records Bureau develops independent school norms for the tests used by its member schools. Many test specialists believe that such specialized test norms as these have greater utility than general norms. Both types, no doubt, have their place.

4. *CEEB norms.* The College Entrance Examination Board tests illustrate still another type of norm. The CEEB norms (mean of 500, range from 200 to 800) were based on the per-

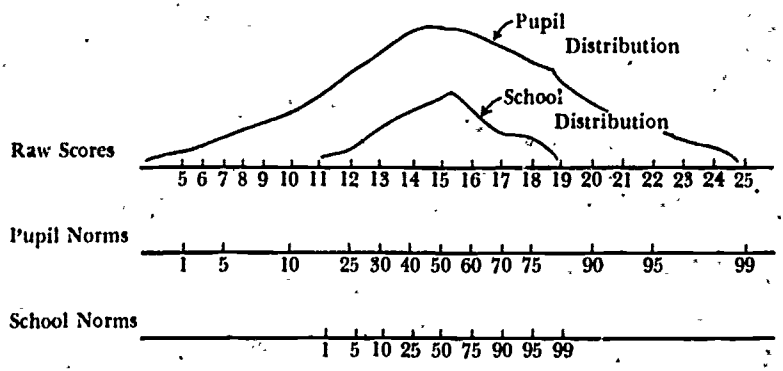
formance of those college applicants who took the Board tests in 1941. Subsequent Board tests have been equated back to the 1941 group. The reader may well question the adequacy of norms which go back more than 20 years. There certainly is reason to assume that college-bound students now are not the same as college-bound students in 1941. However, the practice can be justified simply by remembering that norms may be thought of as a yardstick. Anthropologists tell us that children are taller now than they were a generation ago. Yet we measure them with the same yardstick. In like fashion, college-bound students in 1964 may be more or less able as a group than those taking the Boards in 1941. Yet, within any year's group of college-bound students, they will have the same relative position in relation to each other regardless of the norms used. They may or may not have the same mean of 500, but that is not particularly important. Another way to look at this point is to realize that the actual norms of the College Board tests are not as important as the range of scores and average score of a particular college to which an applicant is applying. For example, if Jim has a V (verbal) score of 480 and an M (mathematical) score of 540, it does not really matter that one is slightly above 500 and one slightly below. What really matters is that Jim has applied for admission to two different colleges. Ivy College has a student body with Board scores running from 350 to 650 and with a mean of 450; State College has a student body with Board scores running from 450 to 750 with a mean of 600. Jim might be accepted by either school, by one, or by neither. If accepted by both, he could choose to enter Ivy, where he would be slightly above average in general ability; or State, where he would be slightly below average. In either case, the important comparison is with the norms of the colleges to which Jim applied, not the 1941 yardstick that was used to get the measure of ability.

5. *School norms vs. pupil norms.* One of the most puzzling aspects of test norms to many people is the difference between school norms and pupil norms. Up to this point this brochure

has been concerned with pupil norms. Looking back at the XYZ Arithmetic Test, we hypothesized that 300 school systems were chosen in the sample but only 240 systems participated. The norms presented in Table 1 were based on the imaginary sample of 20,000 8th graders, with all 20,000 scores put together in one distribution. That process produces pupil norms.

Now let us suppose that we had done something else. For each of the 240 school systems that participated in the norming of the XYZ Test it would be possible to obtain a single mean (average) score. Thus, we could get 240 mean scores. The highest mean score for the school with the most able student body would not be 25 or 24 or 23 or any other score so close to perfection. It might be that some one or two schools would get mean scores as high as, say, 19, and possibly some others as low as 11. However, most of the 240 mean scores would fall at about 14, 15, or 16. There would, in fact, be a distribution of scores, 240 of them, forming a fairly normal distribution with 15 in all probability as the middle of the distribution. Such a distribution of scores would serve as a basis for school norms rather than pupil norms. Figure 1 shows the difference.

Figure 1
XYZ Arithmetic Test
Pupil Norms and School Norms



Now let's look at two specific examples:

Horace Mann Community Schools–mean XYZ score for all pupils, 17

Pestalozzi Area Schools–mean XYZ score for all pupils, 12

In the case of the Horace Mann schools an average raw score of 17 would be a school norm of the 90th percentile. That is to say, the pupils in the Horace Mann schools, *as a group,* achieved at an average level higher than 9 out of 10 of the 240 *school systems* in the XYZ norm group. But, that does *not* say that the average or typical pupil in the Horace Mann schools scored above 9 out of 10 *pupils* elsewhere. Actually the average pupil at Horace Mann scored at the 70th percentile on pupil norms (see Figure 1), or above 70 per cent of a group of typical pupils elsewhere.

In the case of the Pestalozzi schools the average raw score of 12 can rightfully be interpreted as a school percentile of 5. Only 5 per cent of a group of 100 typical American *schools* could be expected to achieve average scores below the Pestalozzi schools (if we consider our sample of 240 schools as typical). However, the average pupil at Pestalozzi scored above 25 per cent of a group of typical pupils and below 75 per cent (see Figure 1).

Both school norms and pupil norms serve useful purposes. But the purposes are different and should not be confused, one with the other.

6. *Comparable forms.* Many standardized tests are published with more than one form. The different forms may be referred to as comparable forms, parallel forms, or equivalent forms. Each form of a test is designed to be as nearly like the other form(s) as it is possible to make it through careful item selection. If all forms of a test were exactly alike, if they correlated +1.00 and had the same mean and standard deviation, then one could use the same norms for all forms and expect them to be entirely satisfactory. However, to the extent that the correlation between two forms of a test is less than +1.00, and to the extent that the means and standard deviations vary a bit from form to form,

a question arises as to the appropriateness of using identical norm tables. Of course, if the correlation is +.99 and the means and standard deviations vary by only a tenth of a point, one certainly would not question the use of the same norms. But if the correlation between two forms is only +.80, with means and standard deviations varying by several points, one would certainly feel that each form should have its own separate norms. Some place there is a breaking point beyond which it is not appropriate to use the same norms for two forms of a test. That breaking point cannot be flatly specified. In practice some test authors will supply differentiated norms when the differences are very small; other authors will tolerate larger differences under the same norms.

When a test provides different norm tables for Forms A and B, the authors are saying, in effect, "In spite of the fact that we attempted to build equivalent forms there still is enough difference between the two so that it is essential to provide separate norms." When another test provides one set of norms for Forms X and Y, the authors are saying, in effect, "The two forms are so close in their statistical properties that it is not worth while to provide separate norms." The reader should realize that the fact that different authors take these different points of view *does not* necessarily mean that the Forms X and Y actually are closer to each other than Forms A and B. It may just mean that the different authors have used different standards of statistical rigor in making their decisions. For example, the comparable form reliability between A and B may be .85 and between X and Y also .85. Then the author of Forms A and B has said, "A correlation of .85 is not high enough to warrant the use of the same norms", while the author of Forms X and Y has said, "A correlation of .85 is high enough to justify the use of the same norm tables."

All of this leaves the consumer—the teacher, counselor, administrator—in a somewhat awkward position. Unless he plans on developing his own local norms for every form of a standardized test that he uses, he must accept the decision of "same" or

"separate" norms that the test author has made in developing the test. He can keep in mind, however, that the development and use of separate norms for each form of a test is, statistically, the conservative approach. It makes no assumption of complete or almost complete identity of the two or more forms of a test. Rather it treats each form independently and establishes norms for each. Thus, it is the more accurate of the two approaches.

7. *Selection scores decrease the need for test norms.* In some situations the selection or "cut-off" score is more important than the norms. Thus, in a given college, the norms of the CEEB tests may be relatively unimportant because the college has decided upon the range of scores it will accept.

We can see the same thing in a simple example. Suppose a school has used an algebra aptitude test for a number of years. The administrator probably has a pretty good idea as to the meaning of certain raw scores, without reference to any norm table. For example, he may have noticed that not a single student with a raw test score below 20 has ever passed algebra in his school, and that no one with a raw score below 23 has ever got a C or better. It does not really matter to him whether a raw score of 20 is equivalent to a national percentile of 5 or 25 or even 50 or 75. What matters is that he knows what certain scores imply for success or failure in algebra as it is taught in his school by his teachers to his pupils.

Another type of situation in which norms are not very important occurs when a limited number of pupils are to be selected out of a group. Suppose a high school is instituting a new Advanced Placement course in physics. It has 30 pupils who request enrollment in the course, but enrollment is to be limited to 15. If ability test results are to be used as a part of the selection process, it does not matter much what the actual IQ's (or percentile ranks) are. It probably is more important simply to put pupils in rank order in terms of ability test scores. For that purpose, raw scores would work just as well as IQ's or percentiles.

There are not many school situations in which selection scores are all-important and norms of no importance. Generally, school administrators and counselors are concerned with multiple criteria for selection. They wisely include past grades, teachers' recommendations, and evidence of motivation as well as test scores, so that any one test score is just another bit of evidence. Nevertheless, in some school situations the relationships of test scores to demonstrated success or failure are more important than test norms.

Avoiding Confusion with Summary Statistics

Occasionally the word norm is given a completely different meaning than it generally carries, being used to refer to an average score (to a mean or to a median). When a principal asks his director of testing: "What is our norm on the XYZ Test?" he probably wants to know the school average, not a listing of raw scores or percentiles such as that in Table 1. This adds a meaning to the word norm that it would be better to ascribe only to mean or median. Since there are precisely defined words to use for the different measures of central tendency, there is no reason to add another term. Even the word average is more acceptable to describe a measure of central tendency than the word norm. Still, it is wise to keep in mind that this other, additional meaning is used rather widely. The educator who wants to appear knowledgeable in the test domain should avoid using the word norm when he wants to refer to an average score, just as he should avoid using the word correlation when he is speaking simply of a relationship between two things.

A Special Question: Whether or Not to Develop Local Norms

As has been mentioned previously, the only basic difference between local norms and national norms is in the defined population. Statistically there is no difference in the way one computes local norms. However, in the total process of developing norms there are some steps that may be left out when developing local norms. In all probability local norms will be developed after a test is already in use in a school system. Therefore there are not going to be any sampling problems. If the XYZ Arithmetic Test is adopted by a particular school system, all 8th graders—not just a sample of 8th graders—will be given the test.

Characteristics of Local Norms

In a middle-sized system it is often advisable to use the scores of all pupils taking the test in a particular year to establish local norms. In very small systems, scores may be combined over several years to get the norms. In larger systems it is not necessary to use all the test scores. Thus, one might use only one-half or one-fourth or one-tenth, or some other appropriate fraction, making sure that the portion used is selected randomly, and that at least several hundred scores are used.

There are two special points to keep in mind when deciding whether to develop local test norms. They are: 1) local test norms do *not* change the rank order position of any pupils from their positions on national norms; and 2) local test norms simply move a pupil's score up or down (or leave it unchanged) from national norms.

The position, or rank order, of any pupil's score is determined by his raw score, not by a norm. A normative score is simply a more convenient score for expressing that ordinal position. If Joe got 20 items correct on the XYZ Test and Steve got only 14, Joe will have a higher percentile (or stanine or grade equivalent, etc.) on national norms or regional norms or state norms or local norms or building norms, or any other norms one cares to develop. When put this way the statement is so obvious that one may question the need for even saying it. However, too many users of tests feel that a local norm somehow provides different information about a pupil than a national norm. Actually local norms provide the same information, but they relate that information to a different group.

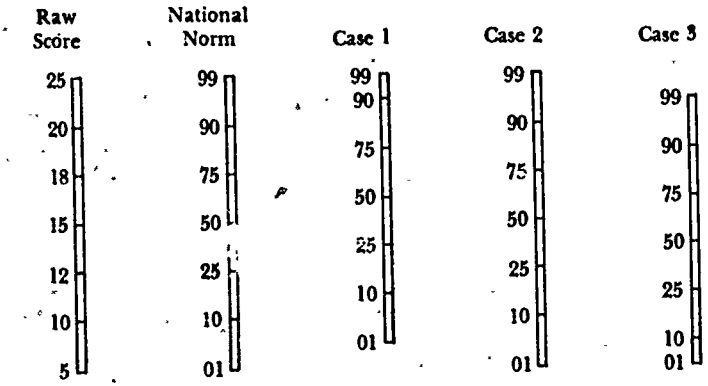
That brings us to the second point. If we refer back to the idea of test norms as yardsticks, we can say that local norms provide us with a yardstick that has different units (or numbers) on it than those on the national norm yardstick. In practice, then, this boils down to the following relationships between national and local norms:

1. If a local school system has a student body that tends to score above national norms, its own local norms will lower almost everyone's derived score below what it would be on the national norms.
2. If a local school system has a student body that tends to score very close to national norms, its own local norms will not be materially different from national norms.
3. If a local school system has a student body that tends to score below national norms, its own local norms will raise

almost everyone's derived score above what it would be on the national norms.

Table 2, based on the XYZ Test, illustrates the differences in national and local norms for the situations described.

Table 2
XYZ Arithmetic Test
National and Local Norms



Think of a pupil, Jane, who got a raw score of 15 on the XYZ Test. On national norms she would receive a percentile rank of 50. If she attended a school system with an above average student body (Case 1), her percentile rank on local norms might be about 35 or 40. If she attended a school system with an average student body (Case 2), her percentile rank on local norms would be close to 50. If she attended a school system with a below average student body (Case 3), her percentile rank on local norms might be about 65 or 70.

To interpret these situations one would say, in Case 1, Jane scored above half of a typical group of 8th graders nationally on the XYZ Test, but scored above only 35 per cent of the 8th graders in her own school system. For Case 3 one would say, Jane scored above only half of a typical group of 8th graders nationally on the XYZ Test, but above 65 per cent of the 8th graders in her own school system.

In neither case did Jane's test performance change. She was average on a national scale. The only things that changed were the two different groups (yardsticks) against which her score was compared on a local basis.

Pros and Cons of Local Norms

There are differences of opinion as to the value of local norms. Proponents feel that they are more useful to a school system than national norms; detractors feel that little is gained by the efforts necessary to develop them.

Perhaps the most important point made by proponents of their use is that local norms provide a fairer assessment of local competition. That is, if Jane is in a school such as Case 1, it is better to know that she is a bit below average in relation to her classmates (35th percentile) in arithmetic than to know that she is average in relation to a national group. Or, if she is in a Case 3 school system, it is better to know that she is above average in arithmetic (65th percentile) in relation to her classmates.

Opponents of the use of local norms might say that, while it is well and good to compare Jane to her classmates, she may not always be in that school system, and that it is better to think of her arithmetic achievement in relation to pupils all over the country. They might say that it is best to think of Jane as having average proficiency in arithmetic, not below average just because her classmates happen to be particularly able in arithmetic, or above average just because her classmates are low in arithmetic achievement.

Fortunately this question of the greater value of local or national norms is one that does not have to be resolved. There is no reason why a school system cannot use both types of norms (as well as any others that are appropriate) and thus gain the advantages of both of them.

As far as Jane's junior high teachers are concerned, it may be of more interest to know where Jane stands in relation to her

classmates than in relation to a national sample. Considering the day-to-day competition that Jane faces in her classes, this is a reasonable view. However, Jane's counselor, while he certainly will be concerned with her achievement within her own school, must also look beyond the local situation to the potential competition Jane may face in other schools or after her school career is ended. For example, Jane might be enrolled in a junior high school with a below average student body (so she appears above average) but be headed toward a senior high school with an average or even above average student body. In such a case the counselor could use national norms as a base (or yardstick) that will not change as Jane's school changes. Of course he also could use both the junior high and senior high norms to complete the picture of present and future competition.

Special Considerations

Before leaving the topic of local norms there are several important points to be made. The first is simply a plea for common sense when deciding whether to develop local norms or not. There is nothing in the process itself that would preclude a school system from developing local test norms for a group intelligence test. But the very idea of intelligence as a general mental ability is foreign to the idea of establishing separate IQ tables for each school sysem. The idea is so foreign that the author knows of only a few systems that have ever done it—and they were very large systems which felt they could demonstrate that their local norms were comparable to national norms anyway. This extreme, and perhaps ridiculous, example is given to emphasize that while local norms have a place, that place does not encompass all types of tests.

The best case for developing and using local norms can be made with respect to achievement tests. The goals or purposes of achievement tests are closely related to the specific goals of a school system, and the day-to-day operation of most classrooms is geared to increasing the knowledges, skills, and understandings

of the pupils in the system. It makes sense to consider the use of local norms with achievement tests; it does not make sense with intelligence tests or any other test seeking to measure some aspect of human personality not closely related to school systems individually.

The second point to be made here is that any school using local norms on a test is and always will be "average" on those norms. It is impossible for a school to be above average or below average on its own local norms. The process of developing local norms automatically assigns the middle raw score to a percentile of 50 or a stanine of 5 or a T-score of 50—the middle score of whatever type of derived score is being used. When a school person speaks of a system's being above or below average, he must be relating local statistics to national or regional or state norms. Again, one should not confuse local norms with summary statistics.

Salient Considerations in the Interpretation of Norms

Several years ago the author was approached by an elementary school principal who wanted some help with the standardized testing program in his school. In the course of the conversation the principal expressed concern over the reading level of his 5th grade pupils, because, "Forty percent of the fifth graders are reading below grade level on a standardized test of reading." He was rather nonplussed when the author congratulated him on the apparently good job of reading instruction going on in his school (a school with average, not above average pupils). If the reader, also, is puzzled let him think back over the meaning of national norms and the way they are developed. On national norms half (50 per cent) of all pupils are reading at or below grade level. By their very nature, norms automatically pick the middle score of a distribution and define that point as average, as grade level, as 50th percentile, or as the midpoint of whatever scale is being used. Thus, a school with only 40 per cent of its pupils below grade level (rather than 50 per cent) is above average (assuming that most of the other 60 per cent are above grade level).

In this example the principal made the very common error of translating test norms into standards of achievement. He had assumed that all or almost all pupils should be reading *at* grade level. But grade level on test norms is simply a point (or score) dividing all the pupils in a grade into halves. One may or may not like the way test norms are defined and developed, but it is hardly proper to interpret them as if they were something that they are not. Suggestions have been made from time to time that we need standards of achievement in such skill areas as reading, arithmetic, and language, so that we could compare individual or group achievement with those standards. Such standards might be very useful, but attempting to change norms into such standards is an impossible task.

Individual Pupil Interpretation

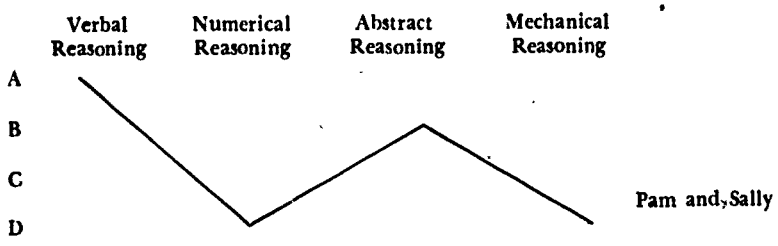
This brochure is not directly concerned with all of the intricacies of test interpretation. It is primarily concerned with test norms. But, since one of the purposes of test norms is to develop and communicate meanings and understandings about individual pupils, test interpretation cannot be ignored. Our purpose here will be simply to point up those characteristics or aspects of test norms themselves that must be kept in mind, and not to cover test interpretation in any great depth.

There are two general ways of using norms in test interpretation. The first is to use norms as a yardstick for comparing a pupil with himself, for comparing his own high points with his own low points, or for comparing his performance over time. If this is the chief point of interest, then the norms help us see in which areas a pupil scores high and in which he scores low.

Part I of Figure 2 shows the profiles of two girls, Pam and Sally, who took an aptitude test having four parts, verbal, numerical, abstract, and mechanical. The letters A, B, C, and D are neither raw scores nor norms. They simply designate high and low performance for each pupil in relation to her own perform-

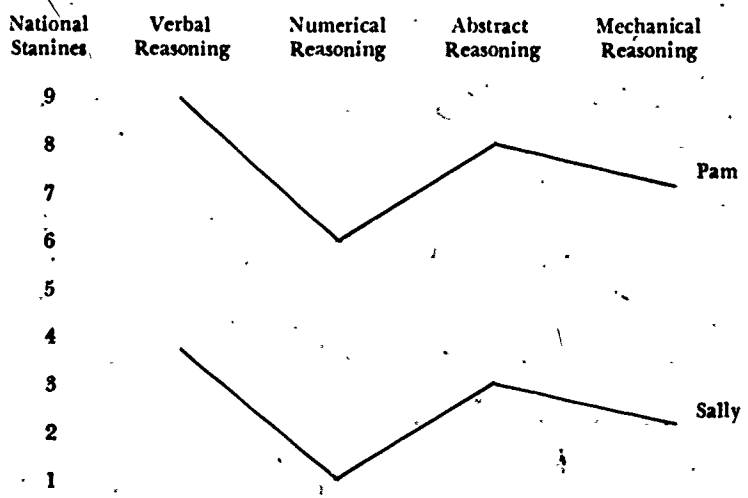
ance on these tests. Thus, both girls got their highest score on verbal; both got their lowest score on numerical (exactly three units below verbal); both had abstract scores one unit below verbal (and two above numerical); and both had mechanical scores two units below verbal (and one above numerical). Thus, with respect to their own strengths and weaknesses on the four aptitudes being measured, Pam and Sally are identical. If a counselor were discussing Pam's or Sally's profile with her, the counselor and counselee might come up with very similar interpretations in both situations: greatest strength in verbal ability, least in numerical, with abstract and mechanical in between. Perhaps that is sufficient, perhaps not. The whole point of this example is to illustrate the extreme situation of evaluating strengths and weaknesses in relation *only* to the individual himself.

Figure 2
Pupil Profiles, Part I



The other way of using norms in individual pupil interpretation is to use them as a yardstick for comparing a pupil with other pupils similar to himself. The question here is not so much where one is high or low but whether one is high or average or low in relation to other pupils. The primary point concerns the relation of the individual to the group, rather than the individual to himself.

Figure 2
Pupil Profiles, Part II



Part II of figure 2 shows that Pam's scores on this aptitude test were all high (from stanine 6 to stanine 9) and Sally's were all low (from stanine 1 to stanine 4). Thus, in relation to a national sample of all pupils of the same age and grade level, Pam scored very high on verbal reasoning, definitely above average on abstract and mechanical, and high-average on numerical; Sally scored low-average on verbal, definitely below average on abstract and mechanical, and very low on numerical. This interpretation says very different things about Pam and Sally, whereas the interpretation in Part I said the same thing for both.

Both interpretations are correct. They differ simply because the yardsticks used were different. Pam and Sally have performed only once on this test.

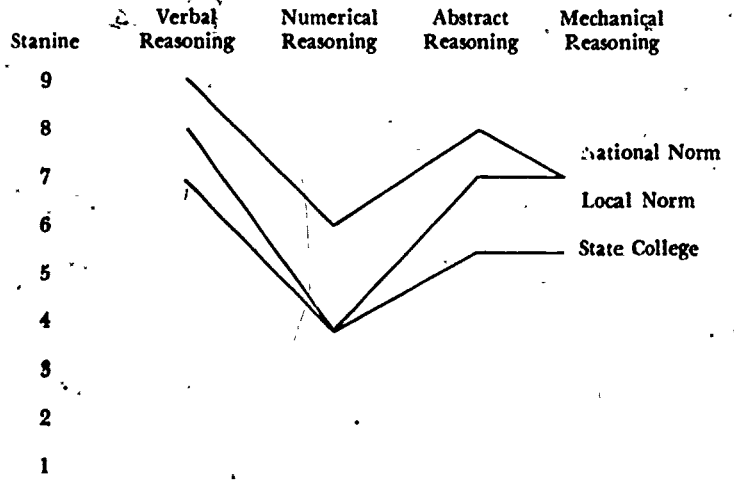
These two extreme examples were drawn in order to point up individual test interpretation in relation to self and in relation to peers. Of course, there is no reason why a counselor or teacher or administrator cannot follow *both* courses, looking both for individual strengths and weaknesses and for the general ability or achievement level in relation to other pupils.

The Case for Multiple Norms

In the discussion of local norms the point was made that there is no reason why a school should use only local or only national norms or only any other kind of norm. Each different kind of norm that is used has the potential of adding some bit of meaning to a total evaluation.

Let's look again at Pam's scores on the four-part aptitude test that we explored in the previous section (see Figure 3). Pam had stanines of 9, 6, 8, and 7 on the national norms. Let's assume that the director of testing in her school system also had developed local norms to judge the competition within the school. On the local norms her stanine scores were 8, 4, 7, and 7. In general, then, her derived scores were a bit lower on local norms, which says that the school Pam attends has above average pupils on the aptitudes being measured, with the possible exception of the area of Mechanical Reasoning. Also, suppose that Pam's parents had attended State College and that she and they are interested in how her abilities compare with students attending State. Fortunately Pam's counselor has been able to obtain in-

Figure 3
Three Kinds of Norms



formation from State with their own norms for this aptitude test. From that information he can tell Pam that her stanine scores in relation to State freshmen are 7, 4, 5, and 5.

There is value to be gained from each one of the different sets of norms. A very strong case can be made for the use of multiple norms—for forcing pupils and parents as well as counselors and teachers to look at test results from two or three or four different points of view rather than letting them see only a single percentile, a single stanine, or a single grade placement score.

Group Interpretation

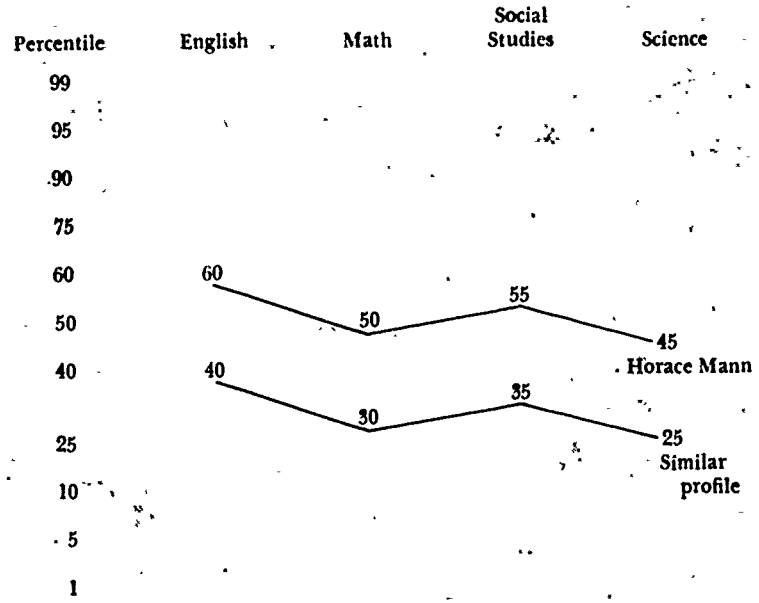
The general principles of individual test interpretation apply also to group interpretation. One can analyze group results obtained within a single school system, and at the same time compare group results from one system to national or regional or state results.

When one speaks of comparing group results, one generally is talking about comparing one aspect of group results (one point in the distribution). The most common thing is to compare medians or means. Sometimes other points, such as the quartile points (25th and 75th percentiles) and the highest and lowest scores also are compared. In any event a single point or a limited number of points in each distribution are compared.

Suppose that the 11th graders in the high school of the Horace Mann School System produce the results in Figure 4. Their average (median) percentiles are 60 in English, 50 in mathematics, 55 in social studies, and 45 in science. The first type of comparison causes one to see that their highest score is in English, the next is in social studies, the next in mathematics, and the lowest in science. In this type of comparison one is concerned only with rank order of performance within the single school system. Having established the rank order, one begins to question why it came out as it did. Does it simply represent chance, with differences so small as to be insignificant? Or are the differences real? If so, why do Horace Mann pupils do their

best work in English and poorest in science? One can determine statistically whether the differences are significant or not. If they are, one then must rely upon the Horace Mann principal and teachers and curriculum director to "explain" the differences. In this example the same rank order would have appeared if the percentiles had been 40 and 30 and 35 and 25 respectively (see Figure 4). The rank order is the crucial point.

Figure 4
A Group Profile



In the second type of comparison one is concerned with the arithmetic size of the medians. Thus, one could say that, when compared to a national sample, Horace Mann pupils are average or above in English, mathematics, and social studies and below average in science. Or, it might be more reasonable to think of average as a range, not a point, and say that Horace Mann pupils were about average in mathematics, social studies, and science and were above average in English. In this comparison, a group comparison, a smaller variation from the median may be considered significant than when working with individual results.

When using norms for group comparisons of any sort there are several points to keep in mind. One of these is that in comparing achievement test results with national norms it is important to consider possible group differences in scholastic aptitude or general intelligence. In our Horace Mann example we saw that the school appeared average in three areas and above average in one. This statement assumes that the pupils in Horace Mann are of average ability. If they were of above average ability, their group achievement would not look so good; if they were of below average ability, their group achievement would appear very fine indeed. Average ability level cannot be ignored when comparing group results with an external norm (national, state). Average ability is of less importance if one is making high and low comparisons within one school.

Another very important point in the area of group evaluation is that the total group or a random sample must be tested if group comparisons are to be made. It is not reasonable to test only part of a group and then draw inferences that apply to the total group. The most obvious deviation from this principle in recent years has been in relation to inferences some people have drawn from results on the *National Merit Scholarship Qualifying Test* (NMSQT). The NMSQT is a test of general educational development, a type well suited to general curriculum evaluation. The authors have developed national norms for the test by relating it to another widely used test. Thus, it is possible to use the NMSQT for group evaluation *if* a school tests an entire grade or a representative group with it. However, most schools use a self-selection process, with many college-bound pupils and a few others taking the NMSQT. Such a sample is not at all typical of the total grade, and the group results from such a sample are almost meaningless. There is no known reference group against which one can compare results. Even the norms built only on those pupils taking the NMSQT have very little meaning, because of the impossibility of defining the population exactly. Some schools test only a handful of pupils with

the NMSQT, some a sizable group, and some everyone. The mixture makes comparisons impossible.

Another point worth mentioning in relation to group norms is the use of summary statistics on group results for publicity purposes. Some systems take pride in letting their communities know that group test results are well above national norms. There is certainly nothing wrong in being proud of a job well done, providing the evidence actually says that the job has been well done. But both of the points above should be kept in mind in this connection. Achievement test results at the 70th percentile in English and mathematics and social studies and science are not outstanding if the average IQ also is at the 70th percentile level. Having ten National Merit semifinalists in a large high school with an average IQ of 120 may not be any more laudatory than having one semifinalist in a high school with an average IQ of 90. Again, it is fine to be proud of outstanding group achievement and outstanding individual achievement, but it is unrealistic for a school to assume all of the credit for such achievement.

Special Considerations

In interpreting test norms there are a few more points to be made. They do not apply just to individual use or just to group use. They are not considerations that are systematic and regular; rather they vary from situation to situation. The first of these is not really a characteristic of norms at all; it is a characteristic of our society and particularly of parents in our society; it is a feeling. It is the feeling that somehow being "below average" is a stigma or a curse; that in fact being "average" in our society is not enough. Many parents hope—and really expect—that their children will be above average. When a parent is faced with average or below average grades or test scores for his youngster, there often is a feeling of resentment, a feeling that the school has somehow let him down. The parent is apt to feel that he has provided the school with an above average boy or girl, and

that any failure to maintain that position is bound to be the fault of the school. This tendency on the part of so many adults has important implications for test interpretation.

To satisfy this feeling, what is needed, perhaps, is a "psychological norm" that is low enough so that almost all pupils find themselves above that norm. To be more realistic, what may be needed is a series of objective standards of minimum achievement for various grade levels and/or subject areas that reasonably can be met by most pupils. Until such standards are developed we are left with the situation in which norms, by definition, say that half of all pupils fall at or below the midpoint.

Another consideration for a test user to keep in mind is clearly and specifically a test characteristic. The title of a test does not define the content of the items in that test. Two reading tests designed for use in grade 7 may or may not measure the same reading skills. If one examines several of the widely used reading tests found within junior high achievement batteries, he will find one reading test consisting entirely of test items which relate to short selections that are read while the examinee is taking the test. He will find another reading test that includes some items of that type but also includes items of a study-skills nature within the reading test. Reading scores from two tests that differ systematically in the types of items they use represent performance on two different tasks. They should not be labeled identically nor interpreted identically.

It was mentioned in an earlier section that, despite the work of authors and publishers in developing national norms for a test, some variation creeps into the norms from sampling inadequacies. There are several types of systematic differences that may occur. One of these is the variation that may occur in mean (or median) values from one set of norms to another. It might seem that such differences should not appear, since, by definition, the middle score for any set of norms is set at the median position. The reason variations occur is that no two samples, even from the same defined population, will be identical.

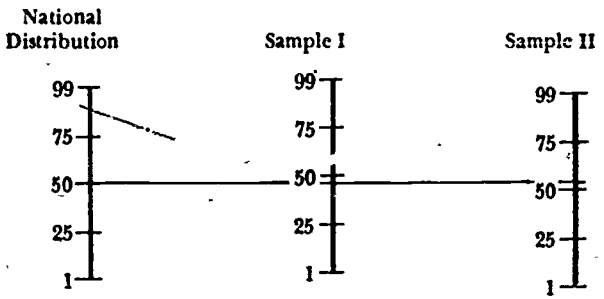
In Figure 5 let's assume that by some magic means we know the exact national distribution of all 10th graders on their knowledge in general science. Testmaker I develops Test I for general science. Testmaker II develops Test II for general science. Both are good tests, but Test II has a slightly larger percentage of physics-type items, while Test I has a slightly larger percentage of biology-type items. Test I is normed on 10,000 pupils in 20 states representing all geographic areas and sizes of schools. Test II is normed on 20,000 pupils in 30 states representing all geographic areas and sizes of schools. None of the pupils tested with Test I was tested also with Test II in the norming process.

Now assume that 10th greaders find biology-type items somewhat easier to do than physics-type items. Also, assume that the 10,000 pupils tested with Test I actually are slightly above average in general science achievement. Putting these two assumptions together we find that the median for the sample of 10,000 is a bit above the "true" median for *all* 10th graders on *all* types of general science items. This is pictured as Sample I in Figure 5.

Again, assume that 10th graders find physics-type items a bit more difficult than biology-type items. Also, assume that the 20,000 pupils who took Test II actually are slightly below average in general science achievement. Putting these two assumptions together we find that the median for the sample of 20,000 is a bit below the "true" median for *all* 10th graders on *all* types of general science items. This is pictured in Sample II in Figure 5.

These conditions lead one to two sets of norms for the two different tests of general science though they are supposedly measuring the same achievement characteristic. Theoretically, if both tests sampled exactly the same skills and both normative samples were identical, the two sets of norms would be identical. In practice they will vary a bit, as in our example. To extend our example to the area of individual interpretation, one can imagine a pupil who is exactly average on our hypothetical

Figure 5
General Science Test



national scale. Such a person would get a percentile rank of perhaps 46 or 47 Test I, and of perhaps 53 or 54 on Test II, depending on how much each sample varied from the "true" situation. In like fashion a "true" percentile of 75 might be a 70 on Test I and 80 on Test II. ("True" is defined here as the exact knowledge or ability that an individual possesses; it cannot be measured.)

Very few formal studies have been made of mean differences in test norms. Test publishers are not in a position to make them, and few other agencies have the resources. The studies that have been made generally are within a single city or state. Certain state testing programs lend themselves to such studies. Since tests are revised and restandardized periodically and each restandardization involves a different sample, there is need for continuing research on the comparability of norms for widely used standardized tests.

In Conclusion

Almost inevitably a discussion like this one focuses on the *problems* in an area—and, in the very process of explaining and trying to simplify things, it may create the impression that those things are difficult and complicated. To be sure, there *are* some complexities in the sensible use and interpretation of test norms, but they are rather modest ones and a person can learn to handle them with ease.

Anyway, even if there are some problems, they are certainly worth wrestling with, because the potential gains are so great. The scholarly, scientific work of the testmakers and the companies engaged in testing has built up a tremendous, unprecedented body of resources for American education. It enables us, as never before, to diagnose difficulties, make thoughtful predictions, and evaluate the successes or failures that grow out of our work.

Surely it is worth real effort to capitalize on the possibilities of such resources. Yet most of their value can be thrown away—in fact, they can lead to damage—if we are not competent and wise in interpreting and using the results. Therefore, we cannot resist concluding with some generalized remarks which go somewhat beyond the technical scope of this booklet.

Despite the fact that tests are growing more accurate and precise, year by year, the wisest educators still use their results with a certain *moderation*. In view of all the possibilities for variation, it is a little unreasonable to act as if a youngster's recorded IQ, his score on a personality inventory, his percentile on an

aptitude test, or even his stanine placement on an achievement test represents *exactly* his true ability, his real characteristics and aptitudes, or his actual accomplishment. The state of his motivation, his health, his happiness, may have driven him unusually low or high on a given day. Maybe on another day—or with another tester—he might have scored quite differently. The test may not even have fit him, or the curriculum of his school, very well. And then, with the small sampling of knowledge in a given test, sheer luck may have run for him or against him. Especially near the middle of the distribution a few items can make a striking difference in percentiles. (Look back to Table 1, for instance; the difference between missing or solving two items is the difference between the 40th and 60th percentiles.) On the whole, it is better to think in rather broader terms: "middle range," "high average," etc. The makers of one famous personality inventory specify that no T-score between 40 and 60 is to be thought of as "different" from the mean.

Furthermore, valuable as test data are, they are only *one* form of evidence. After teachers and counselors and administrators have lived and worked with a child for years—after their intuitions and judgments about him have slowly coalesced—and after he has accumulated a record of successes and failures, as represented by grades—it would be sheer folly to write off all such evidence in favor of one or a few scores on paper-and-pencil tests, no matter how good those tests are. We have not yet developed any formal procedures to replace the judgments of teachers and counselors and administrators; we do have formal procedures to provide evidence that will help to improve those judgments.

None of this is to downgrade the enormous importance of good testing programs. Used *together with* all the other evidence we can get, test data are a marvelous aid to better planning and teaching. To us in the profession the great challenge is to use this new resource perceptively on behalf of every child and youth in our care.

About the Author

When Dr. Frank B. Womer speaks of tests and norms and everything pertaining to the use and interpretation of standardized measures, he does it with an exceptional authority which is the product of his entire training and practical experience. At present he is Associate Professor of Education at the University of Michigan where, among other duties, he teaches courses in achievement and ability testing and test construction. In his years at the university he has taught also in the field of educational psychology, and he has served as Consultant on Testing and Guidance in the Bureau of School Services.

These days he spends much of his time as a consultant to elementary and secondary schools in the area of test program development and the use of test results. He is also Director of the Michigan School Testing Service, and he is responsible for the annual Michigan School Service Testing conference, attended recently by more than seven hundred persons.

After graduating from the University of Colorado in 1948, Dr. Womer taught mathematics for two years at Alamoso High School in that state. Since then he has worked chiefly in Michigan, taking the M.A. from the University of Michigan

in 1951 and the Ph.D. in 1956. From 1954 to 1956 he was Test Consultant and Associate Editor in the Houghton Mifflin Co., Boston. He is a member of Phi Beta Kappa, Phi Delta Kappa, and Phi Kappa Phi; he is active in the American Psychological Association as well as in the American Personnel and Guidance Association and the National Council on Measurement in Education; and he is editor of the Council's *Newsletter*.

Very few men in American education have ever devoted themselves so wholeheartedly to promoting intelligent use of the best that modern test development has made available to us.