EDUCATIONAL MEASUREMENT IN HIGH SCHOOL

The Century Education Series

EDUCATIONAL MEASUREMENT IN HIGH SCHOOL

BY

C. W. ODELL, PH.D.

ASSOCIATE PROFESSOR, COLLEGE OF EDUCATION,
UNIVERSITY OF ILLINOIS

Author of
"Educational Statistics," "Traditional Examinations and New-Type Tests"

NEW YORK
THE CENTURY CO.

 361

PRINTED IN U. S. A.

PREFACE

The chief justification for the appearance of this volume dealing with a relatively small division of the whole field of educational theory and practice which is already covered by two excellent treatises is that progress therein is so rapid that there is much new material worth presenting. In the more than three years since both of the treatises referred to appeared, the number of standardized and near-standardized tests suitable for use in the secondary school has increased at least 50 per cent, perhaps even 100 per cent. Not only has there been this increase in the total number, but in several school subjects and other fields of educational measurement, almost all the tests and scales now available have been published within this period.

This book may be thought of as a companion volume to the writer's *Traditional Examinations and New-Type Tests* in that the two together attempt to cover rather completely the measuring and marking of high-school pupils, the earlier one dealing with tests made by the teacher and this with standardized or commercially available tests. It has been written with both the student preparing to teach and the teacher already in service in mind, and it is hoped that the needs of both have been met. It is intended to be elementary enough for the tyro in educational measurement to understand, and yet critical enough to serve as a safe guide in the actual administration of measuring instruments and use of the results secured from their application.

Since so much has already been done and written along the line dealt with, the writer is under obligations to many previous workers. Chief among these are Professors G. M. Ruch and G. D. Stoddard, whose volume *Tests and Measurements in High School Instruction* supplied many data on reliability and other helps, and Professor P. M. Symonds, whose *Measurement in Secondary Education* was an ever-present aid. Thanks are also due many authors and publishers of tests who responded to requests for numerous items of unpublished information.

C. W. ODELL.

EDITOR'S INTRODUCTION

The value of educational measurement as a means of determining the actual accomplishment of students in given types of academic work, through standardized tests of one kind or another, has been recognized for a number of years. While one does not find as many controversial articles with reference to the value or lack of value of educational measurement as he did a few years ago, the number of school systems making a systematic use of various types of tests is steadily increasing. For years the total annual sale of standardized tests has been enormous, estimates varying from thirty to forty million as the number of copies of such tests sold yearly. However, during the earlier years of the use of various standardized tests, they were confined largely to the fundamental subjects of the elementary grades. For many obvious reasons, the development of suitable types of educational measurement for the secondary grades and for the college has been slow. There are many limitations in the successful construction of standardized tests dealing with subjects of complicated character, and it is altogether probable that there never can be as great reliability in tests in many advanced subjects as has been developed in connection with the fundamental operations in arithmetic, spelling, penmanship, and even such subjects as geography and reading in the elementary grades. However, despite these limitations, there has been a decreasing depreciation of the great value of a systematic use of tests in high school, and a large amount of effort has gone into the formulation of all kinds of tests suitable for use in junior and senior high schools.

This volume has been prepared because of a keen realization on the part of the author of a need for a more adequate guide for high school teachers. The very fact that so many tests have been placed upon the market produces in the mind of the average teacher of any high school a very real sense of bewilderment. This book, with its brief and clear statements concerning the

character and purpose of large numbers of tests in the various high school subjects, should become an invaluable manual for the practical guidance of teachers. It is quite evident that the school system which makes a free and intelligent use of measurement from both the diagnostic and the prognostic point of view is bound to be carrying on a much more successful type of teaching than the school system which ignores almost completely the existence of these tests and their very genuine significance in education. The increasing tendency in our larger high schools toward the organization of classes with large numbers of pupils increases the necessity for a systematic use of tests designed to detect the specific weakness of the individual pupil. Given on the part of the teacher an intelligent grasp of the character of the difficulties met by the pupil, she is enabled, even though she has a large class, to modify her work in the light of this knowledge so that the individual can secure far better value from the instruction, and more successfully complete the specific course.

Just as in warfare every new invention leading to more effective offense is promptly met by a corresponding invention looking toward an effective defense against the new danger, so the increasing necessity for the individual teacher to instruct large numbers of pupils carries with it the necessity for equally energetic measures to prevent the submergence of the individual in the mass and thus to offset the danger of lessening efficiency in the case of individual pupils. It seems that, among the various types of measurement described in this volume, almost any intelligent teacher can select means by which not only her general efficiency as a teacher may be improved, but also her ability to assist the individual who otherwise might be doing unsatisfactory work through a failure on the part of a teacher to realize the nature of the deficiencies.

This volume is submitted to the public with the hope that it will meet a very real need for teachers, not merely as a textbook to be used in professional courses in training schools and colleges, but as a handbook for the teacher in her actual work in the school-room.

CHARLES E. CHADSEY.

CONTENTS

FIGURES

TABLES

EDUCATIONAL MEASUREMENT IN HIGH SCHOOL

EDUCATIONAL MEASUREMENT IN HIGH SCHOOL

CHAPTER I

THE PLACE OF MEASUREMENT IN EDUCATION

Purpose of this book.—Ed cation apparently cannot be carried on without a great deal of measurement. Tests and examinations of some sort or other have been in use since time immemorial. Moreover, the tendency of the present age is to increase rather than to decrease the emphasis upon measurement in education, also in many other fields of human activity. Pupils' capacities, characteristics and traits, as well as their actual performances, teachers' efficiency, the condition, appropriateness and completeness of school buildings and equipment, the worth of textbooks, educational costs, and many other factors are being continually measured both formally and informally. Since this is true, it is highly desirable that these measurements be made as accurate and significant as possible. They are as important as measurements of distance, size, or weight, of volume of business, wage levels or labor turnover, of birth and death rates, or of any of the innumerable other things which are measured in the activities of the home, of business, industry, government, and so forth. The purpose of this book is, therefore, to treat of certain types of measurements and measuring instruments that have a close connection with the secondary schools and of the uses that may be made of the results therefrom. Chief attention will be given to the measurement of capacity and performance in school subjects, but the measurement of intelligence, of personality and character, and so forth will be considered as well.

The purposes of educational tests.[1]—High-school pupils and their traits and performances are rated for a number of purposes. The following list thereof is not intended to be exhaustive, but rather to include under a few general heads the more common and significant purposes:

1. Determining pupil classification, including promotion and failure.
2. Stimulating pupils to study.
3. Diagnosing pupils so that they may receive more efficient instruction.
4. Reporting to pupils, parents, and others interested the quantity and quality of work being done by pupils.
5. Serving as the basis for educational and vocational guidance.

There is some overlapping between the purposes named above, but the chief emphases differ. Each might easily be subdivided into a number of minor purposes. For example, the second, that of motivating study, may be accomplished through the use of measurements for any one of a number of means of motivation, such as:

A. Determining the award of honors for which pupils strive.
B. Promoting competition between individual pupils or groups of pupils.
C. Encouraging pupils by indicating that they are making satisfactory progress.
D. Holding before pupils the fact that they are in danger of failing.

It should be remembered that measurements, which in most cases may be interpreted to mean school marks, are not only used for the variety of purposes indicated above, but that they are employed for these purposes in many situations and in connection with many persons and groups of persons. Pupils' marks or other scores are of interest to pupils themselves, to their parents, other relatives and friends, to teachers, supervisors, and administrators, to officials of other schools which pupils are entering or considering entering, to possible employers, and to various other persons. Indeed, this circle is so large that it may be said to include the whole public. Therefore it is evident that

[1] The expression "educational tests" will be used to refer to all tests of subject-matter, intelligence, personality, habits or any other abilities, performances or characteristics of pupils, employed in connection with school work. It is sometimes used in a narrower sense to designate only those tests which measure knowledge of school subject-matter.

not only the public schools themselves but also the system of measurement which they employ is a matter of general interest and concern, and that the making of educational measurements as satisfactory and as nearly perfect as possible is a matter deserving much attention and study.

The general need for more accurate measurements.—The statement has often been made that progress in physical and biological sciences has been made possible in large measure by the relatively exact measurements which can now be made. It is pointed out that progress in any field is largely limited by the degree to which accurate measurement therein is possible. For example, it would be impossible to determine with any considerable degree of accuracy the value of a particular article of food unless the results as shown in increased weight, greater strength or vitality, better health, or some other way, could be measured with some exactness. Similarly the best method of constructing automobile tires could not be determined unless there were trustworthy methods of measuring their life, their riding qualities, and other desirable characteristics. It is, of course, true that some progress could have been made in either of these lines or in almost any other without exact measurements, but in most instances it would have been comparatively small. Thus, to return to the first example just given, if the only means of measuring the value of a certain article of diet were some one's opinion of how an individual's appearance after he had used it for some time compared with his appearance when he began to use it, the science of dietetics would have made a great deal less progress than it has. It is generally agreed that education has advanced considerably during the centuries, but practically no one denies that there is still plenty of opportunity for improvement ahead, and experience indicates that in education as in other fields the more exact our measurements are, the more surely will we be able to make such improvement.

The subjectivity of ordinary school marks.—The traditional method of measuring or marking in education has been by the use of examinations of some sort or other, plus the evaluation of pupils' recitations and other oral work, their notebooks, compositions and other written work, and such laboratory or "life-situation" work as was carried on. It has been customary in the

public schools of this country to rate or mark pupils' work in terms of per cents, although letters and sometimes other symbols have been employed. Despite attempts to define the significance of the percentile or other symbols employed, the fact remains that fixed meanings are rarely if ever attached thereto. Dearborn,[2] Johnson,[3] Kelly,[4] Starch and Elliott,[5] and others have made studies which show that practically all ordinary school marks given pupils' work are rather highly subjective; that is, they depend to a considerable degree upon the person giving them and vary greatly if given by different individuals or by the same individual at different times.

To indicate more definitely the character of the evidence obtained, several examples may be given. Perhaps the most striking findings were those of Starch and Elliott with regard to a geometry paper. The marks given by 116 teachers of mathematics, on the basis of a passing mark of 75 per cent, ranged from 28 to 92 per cent. The complete distribution is shown by Figure 1. From this it can be seen that two of the 116 teachers assigned marks of from 90 to 94 per cent, seven of from 85 to 89 per cent, and so on down to one who gave a mark between 25 and 29 per cent. A second example may be taken from Wood,[6] who reports a study of college entrance papers in algebra and geometry that were scored independently by two different readers. The results show that if about 30 per cent of the candidates are failed by each reader, the chances are that less than 60 per cent

[2] Dearborn, W. F. "School and University Grades," *Bulletin of the University of Wisconsin*, No. 368, High School Series, No. 9. Madison: University of Wisconsin, June, 1910. 59 p.

[3] Johnson, F. W. "A Study of High School Grades," *School Review*, 19:13–24, January, 1911.

[4] Kelly, F. J. "Teachers' Marks, Their Variability and Standardization," *Teachers College, Columbia University, Contributions to Education*, No. 66. New York: Bureau of Publications, Teachers College, Columbia University, 1914. 139 p.

[5] Starch, Daniel and Elliott, E. C. "Reliability of Grading Work in History," *School Review*, 21: 676–81, December, 1913.

———— "Reliability of Grading Work in Mathematics," *School Review*, 21: 254–59, April, 1913.

———— "Reliability of the Grading of High School Work in English," *School Review*, 20: 442–57, September, 1912.

[6] Wood, B. D. *Measurement in Higher Education*. Yonkers, New York: World Book Company, 1923, p. 124–25.

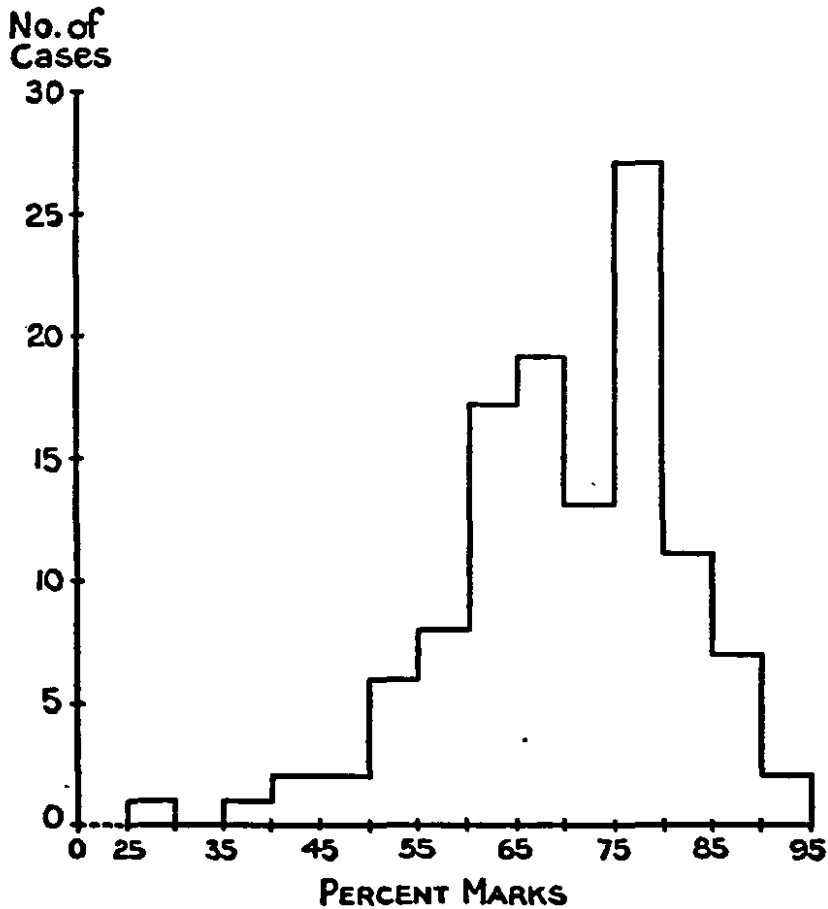


FIGURE 1

DISTRIBUTION OF MARKS GIVEN A GEOMETRY PAPER BY 116 TEACHERS

of those failed by one reader will be the same as those failed by the other. In other words, of those failed by either reader, more than 40 per cent were passed by the other one. Perhaps the following incident, likewise related by Wood,[1] is an even better

[1] Wood, B. D. "The Measurement of College Work," *Educational Administration and Supervision*, 7: 301–34, September, 1921.

illustration. One of a group of college professors of history reading entrance examination papers in that subject prepared for his own convenience a set of what he considered model answers. By some accident or other this paper was rated by several others of the group of readers who thought it a bona fide paper prepared by a candidate. The marks given it ranged from 40 to 90 per cent, although evidently the man who prepared it considered it worth 100 per cent.

Such wide variability as is indicated by the results quoted and numerous other studies not mentioned here is not merely found among different schools, but also among different departments in the same school and different teachers in the same department, and, to a somewhat smaller degree, among the marks given by the same teacher at different times. It exists both with regard to the marks assigned the same specimens of pupils' work by different teachers or by the same teacher at different times and with regard to the total distributions of marks given by different teachers, departments and schools.

Since the most important single basis of determining school marks has been the written examination, it has been largely with the scoring of examination papers and the assigning of marks based on the results therefrom that the studies referred to have been concerned. Not only has it been shown that the marks assigned are decidedly inaccurate, but also that many, if not most, examinations are given without any clear conception of their function. Furthermore, they have in many cases been decidedly laborious for pupils taking them and for teachers scoring them. Hence thoughtful teachers and others engaged in educational work felt the need for better measuring instruments and were ready to give attention, frequently too uncritical attention, to any proposed substitute for ordinary written examinations and tests.

Causes of the subjectivity[8] of ordinary school marks.—Many reasons why school marks as ordinarily given exhibit such great variability as has been found may be stated. Probably the

[8] School marks or other measurements are said to be subjective when they are affected by the opinions, attitudes, personalities, or other sources of bias of the persons who determine them.

chief of these is that different teachers have in mind different bases of marking. Some endeavor to base their marks on the actual performances of pupils. Others take into consideration pupils' capacities and attempt to rate performances at least partially with regard to ability to perform. Others consider the effort apparently put forth by pupils, their attitude toward the subject, and frequently toward the teacher herself, their general behavior, and other factors. Furthermore, teachers disagree as to the phases or types of pupils' performances upon which to base their marks. Some consider chiefly pupils' ability to give back more or less verbatim material acquired from the textbook, the teacher, or elsewhere; others pay more attention to pupils' initiative, their ability to apply what they have learned, and other such results of instruction. Some teachers give considerable weight to speed, others do not. For example, if two girls in a sewing class produce garments of equal merit, but in different lengths of time, one teacher will give both the same mark, another will give a higher mark to the girl who used the shorter period of time. Still another point of difference is that the quality of grammar, spelling, handwriting, general neatness, and other similar characteristics of pupils' written work is either consciously or unconsciously considered by some teachers in determining the pupils' marks in subjects other than English, whereas others pay no attention to such matters of form.

It is ordinarily difficult to get teachers to agree as to just what are the most important points in any body of subject-matter or, in other words, the points which should be covered by examinations. Even when they agree upon this, they are very unlikely to agree as to the relative importance to be attached to each question or exercise. For example, two or more teachers may, although they are not likely to, agree that a certain list of words should be employed for testing vocabulary knowledge in a foreign-language examination, and a certain passage given to test translation ability, but one will probably think that the same number of points should be counted on the list of words as on the passage, another that the words should count more than the passage, a third that the words should count less, and so on.

Another serious cause of disagreement is that, as is shown by the second reference from Wood,[9] supposedly competent scorers disagree as to what are the correct answers. Also they differ as to whether or not any credit should be allowed for incomplete or partially correct responses. For example, in mathematics one may require all answers to be reduced to the lowest terms before they are considered correct and receive any credit, another may not so require at all, and still a third may allow partial credit if the answer is correct in other respects, but not in lowest terms. Similarly in history, if a date is asked for, one scorer may give no credit unless the date given is absolutely correct, whereas another may give full or partial credit if it is approximately correct. In cases in which pupils are required to express their judgments or opinions as, for example, in response to questions asking what probably would have happened if certain historical events had been otherwise, or as to what would be the best course of action in a given situation, opinions of those scoring the responses commonly differ markedly.

Another potent cause of subjectivity and variability in marking is the differences in the attitudes and mental sets of teachers. Some teachers believe that the best way to stimulate pupils to do good work is to be severe in their marking, and thus hold before the pupils the fear of low marks or failure, whereas others believe that the same end is best obtained by encouraging pupils and giving them relatively high marks in the hope that they will thus not be discouraged, but given a favorable attitude toward the work of the class. Some teachers believe that it is an important function of the school to be selective, to eliminate pupils who are not qualified to go ahead and, therefore, fail a considerable number, whereas others believe that the school should rather endeavor to retain all pupils in attendance as long as possible and, therefore, give many high marks in order to encourage pupils to remain in school. Furthermore, teachers are often unconsciously influenced by their likes or dislikes for particular pupils, by pupils' apparent alertness and interest, the neatness of their dress, and other similar factors which tend to produce either more or less favorable opinions of the various pupils.

[9] See p. 7.

Finally, in assigning marks as in any other exercise of judgment or opinion, individuals differ from time to time according to various mental and physical factors. If a teacher is in good health, has nothing that is causing her much worry, and has had no very unpleasant experiences recently, it is probable that she will mark pupils somewhat higher than if the opposite of any or all of these conditions prevails. The writer had one case come under his observation which illustrates this point very forcibly. Five teachers were rating specimens of pupils' handwriting according to the Ayres Handwriting Scale. One of the teachers happened to rate half of the specimens from a given room when she was feeling decidedly ill and discouraged, and the other half some days later when she had completely recovered her health and good spirits. The average difference in the marks given at the two occasions was 20 per cent, although according to the average ratings of the four other teachers there was no appreciable difference in the average quality of the two groups of specimens.

Benefits derived from using objective[10] measurements.—The use of objective measurements rather than those involving personal opinion or bias has many benefits in educational work just as in work along any other line. Many of the possible benefits are not fully realized because the measuring instruments which we are able to construct are not perfectly objective, but only relatively so, nevertheless there are several general advantages which result from the use of such tests as are now available.

Among these one stands out as most important and to a large degree inclusive of the others. This is the increased validity[11]

[10] Objective measurements are those concerning the correctness of which there is no doubt. In other words, an objective test or exercise is one that permits of no disagreement among competent persons as to what is the correct answer. For example, there is no doubt as to the answer to the question, "Who was president of the United States during 1789–97?" or as to the value of x in the equation $2x + 4 = 10$. On the other hand, able and experienced teachers of history will often disagree as to the correct answer to such a question as "What was the most important event during the year 1862?", therefore its scoring is not objective.

[11] A measure or score is said to possess validity when it is actually a measure of whatever it is said or designed to measure. See p. 52 for a more complete discussion.

and reliability[12] of the measures or scores obtained. No argument is needed to support the general principle that if educational measurements are to be made and the results used, they be as exact as possible rather than subject to large errors.

A second distinct advantage is that the unit employed in objective measurements, especially when they are made by means of standardized measuring instruments,[13] is more nearly standard in the sense that it is understood in the same way by all competent persons. In this regard, as in many others, educational measurements are undergoing the same process of development as have physical measurements. Just as at the time when the German foot was an inch and a half longer than the Roman foot there was very liable to be misunderstanding between persons of those two nationalities when they were discussing distances, so at present different persons will not understand the same thing when a pupil is reported as having solved a certain number of problems or spelled a certain number of words correctly. In other words, a problem or a word is not a standard unit, since one problem or word may be many times as difficult as another. This condition is not at all fully remedied by existing educational measuring instruments, but some progress has been made in this direction.

Some of the types of tests suggested in connection with the educational measurements movement enable us to secure much more complete and thorough measures of pupils' ability and traits along certain lines than is possible in the same length of time, or perhaps in any length of time, with the type of examinations almost exclusively employed until the past few years. This is accomplished chiefly by employing exercises of such sorts that the amount of writing to be done by pupils is reduced to a minimum, but also by insuring that pupils understand just what is wanted, by better selection of the points covered by tests, and by other relatively less important means.

[12] A measure or score is reliable if that yielded by a second application of the same or a similar measuring instrument agrees with that which resulted from the first. See also p. 58.

[13] The adjective "standardized" is commonly applied to a measuring instrument which has been carefully or scientifically constructed and employed in a sufficiently large number of cases that one knows what result may be expected.

The amount of time required of pupils for taking tests and of teachers or others in scoring them has been reduced and thus time saved for other important educational uses. The actual marking of traditional [14] examination papers is not highly profitable to teachers, even though something of value may be gained therefrom. If much of the time often used for this purpose were devoted to the more careful construction of tests, to professional reading, to planning instruction, and so forth, the effectiveness of most teachers' work would be increased.

In many cases a desirable change in the attitude of pupils has resulted from employing such tests, since they see that scores thereon are not affected by favoritism of the teacher, misunderstanding of what is wanted, or, indeed, by any factor except possession of the information called for. They can fairly easily verify the scores given their responses and thus have no excuse for claiming that they are not receiving those they deserve.

Objections to objective educational measurements and replies to these objections.—Inasmuch as many attacks have been made upon the modern educational measurement movement and many objections entered against the use of objective or near objective tests, it seems worth while to mention several of the most commonly alleged disadvantages or objections and to reply to each briefly. Eight will be considered, as follows:

1. Objective tests measure only memorized material and knowledge of isolated facts.
2. They are difficult to construct.
3. They do not give training in certain abilities in which traditional examinations afford training.
4. They encourage guessing.
5. They make cheating easier.
6. They confuse pupils as to what they know.
7. They tend to mechanize education.

[14] The term "traditional," "essay," or "discussion" examination is commonly applied to examinations of the sort almost exclusively employed until the last few years and still perhaps more usual than any other. Such an examination consists of exercises that call upon pupils to analyze, discuss, outline, state, summarize, and so forth, that are relatively subjective in their scoring and that require a considerable amount of writing by pupils.

8. The judgments of competent persons are better than the results of tests.

Before answering these in detail, the general reply may be made that upon close analysis most of them will be seen to be objections to the misuse of objective tests and their results, or to unnecessary though sometimes common concomitants of the movement. In many cases arguments against such tests are based on the false supposition that they are intended entirely to displace other varieties of examinations.

1. Perhaps the most common objection to the use of objective tests is that they measure only information or memory and do not measure what are frequently referred to as the higher thought processes, such as reasoning, analyzing, applying, and so forth. There is undoubtedly some truth in this statement, especially as applied to tests as they are rather than as they might be. It is easy to construct objective tests which measure knowledge of isolated facts and mere information, so that too many of the objective tests actually employed do so. On the other hand, it is possible to construct such tests that measure more than this. Furthermore, the reasonable advocates of such tests do not urge that they be entirely substituted for the old type essay examination, but rather that the two types of measuring instruments both be employed, each in its proper place.

2. A second argument against objective tests is that they are difficult to construct, especially in view of the fact that many teachers are not yet familiar with them. This also contains much truth. It usually requires a longer period of time to construct an objective than a traditional examination over the same body of subject matter. This is partly true because by their very nature objective tests tend to require more careful thought in their construction than do essay examinations and, therefore, it is not entirely a disadvantage. The chief answer to this objection, however, is that whatever time is lost in construction will be saved in scoring, at least if the group of pupils tested is as large as twenty-five or thirty. The larger the class the greater the saving in time of scoring. For classes below the size mentioned objective tests may require more of the teacher's time than do discussion examinations. As to teachers' unfamiliarity

with objective tests, this is a condition which is largely being done away with. Furthermore, if it is desirable on other grounds for teachers to acquire such familiarity, the fact that many of them do not have it at present can hardly be considered a convincing argument against the use of such tests.

3. The statement has been made that traditional examinations give training in certain abilities, especially composition and language, not furnished by objective tests. There is no doubt that this may be true, although as the former are frequently administered, with pupils neglecting the form of their responses in the endeavor to put down as much as possible, no desirable results of this sort, but indeed just the opposite, appear to ensue. However, it is not the principal function of examinations to furnish such training, but rather to measure, and their worth should, therefore, be judged chiefly according to their value for this purpose. If secondary advantages can be obtained, well and good, but the primary purpose should not be neglected therefor.

4. Objective tests are said to encourage guessing rather than certainty of knowledge. This may be to a certain extent true of some varieties of such tests, inasmuch as pupils have the chance to select one or more of several suggested answers as the correct one. To this objection there are at least two replies. In the first place this possibility does not exist in all types of objective tests, and in those in which it does exist, its effect can be reduced to a minimum by employing directions that are strongly against guessing and by following methods of scoring that show pupils that in the long run guessing is not profitable. In the second place, situations in which individuals have the opportunity of choosing among several possibilities are very common in life outside the school, and it is desirable, therefore, to have them meet similar conditions in school rather than to endeavor to avoid them.

5. It has been charged that cheating upon objective tests is easier and is, therefore, encouraged by their use. Although it is true that it is often not difficult for a pupil by a hasty glance to discover the single word or other brief response recorded by a nearby pupil, it should not be possible for him to look around enough to secure very much help without being seen by the teacher. Moreover, the same possibility exists in the case of the

traditional examination. Even though a pupil may not be able to look at his neighbor's paper long enough to read a complete answer, he can frequently in a few moments secure important clues.

6. Another charge brought against some varieties of objective tests is that they tend to confuse the pupil as to what he knows by presenting erroneous possible answers. It is undoubtedly true that there is some effect of this sort produced. It appears, however, from both theoretical considerations and experimental results, that if the material dealt with has, as is usually the case, been previously studied and to some extent at least mastered, the danger of confusion is not very great. Indeed, there may even be a distinct gain in fixing points not already known with certainty. Moreover, this also merely reproduces a rather common life situation, the distinction between correct and incorrect when both are present, and there seems to be no satisfactory justification for avoiding it in school when individuals must be prepared to meet it elsewhere.

7. It has sometimes been urged that the use of objective tests, especially those intended for widespread use and sold commercially rather than those prepared by the teacher, will tend to mechanize education. This may be satisfactorily answered upon both theoretical and actual grounds. The use of tests that yield more accurate measures of pupils' performances and capacities would not appear to lessen interest in and attention to the individual any more than improvement in medical diagnosis, for example, has lessened interest in the individual patient. In actual practice the use of objective tests has led to much more emphasis upon proper provisions for individual pupils. Instead of tending to set up a uniform procedure to which all pupils must be made to conform, the modern measurements movement has had exactly the opposite effect. It has revealed and emphasized differences in the capacities, characteristics, and needs of individuals which were scarcely, if at all, realized before. Finally in response to this objection it seems in place to refer to Thorndike's reported statement to the effect that mothers who weigh their babies least often are not those who love them most.

8. It has frequently been urged that the judgments of competent persons are better than the results of tests. If this is true

in educational work, the situation is decidedly unique, as it has been shown that in other lines in which even moderately accurate measuring devices and procedures have been devised they are much more to be relied upon than opinions or judgments no matter how expert. However, we have evidence that no unique condition of this sort holds for education. The school marks dealt with by the studies already referred to were the judgments or estimates of supposedly competent persons who had known the individual pupils being rated for periods of a semester, a year, or even more. If the judgments of a number of supposedly equally competent persons do not agree, there seems little justification for putting much faith in that of any one of them.

Securing observable performances.—It is practically always prerequisite to educational measurement that observable performances be secured from pupils. In some instances such performances are among the natural outcomes of their classroom and studyhall activities. Thus, in drawing, sewing, woodwork, typing, and other subjects, the work of pupils commonly results in concrete objects such as drawings, garments, pieces of furniture, typed sheets, and so on. In other instances the desired outcomes of instruction are not ordinarily observable unless some more or less artificial means of making them so is employed. For example, the desired outcomes from the study of literature are certain items of knowledge, literary appreciation and discrimination, habits of good reading, and so forth, which in general manifest themselves in mental activity that cannot be observed and measured directly. It is necessary, therefore, to have pupils respond to exercises calling for somewhat unnatural but observable responses or performances, ordinarily in written form. Such measurement is indirect, since what is actually scored is rarely if ever just the same as the ability or characteristic which it is desired to measure. Rather, it is a supposed manifestation, expression or application thereof. It is rarely true in mental measurements as it sometimes is in physical measurements, for example, in using the height of a column of mercury to measure temperature, that an absolutely perfect and definitely known relationship exists between the thing actually measured and what it is desired to measure. For this reason, in addition to others, it is important to be explicit in specifying the ability

which is being measured. For example, if pupils are called upon to write a fairly complete account of some historical event within a limited period of time, speed of writing may be a more influential factor in determining the score made than is knowledge of the event named. To give a second example, if pupils are given a list of foreign words with several English words following each from which they are to indicate the one meaning the same, the knowledge of English vocabulary rather than of vocabulary in the foreign language may be measured. If the language dealt with is French, for example, a Frenchman who knew every word in the list but was unable to give any English equivalents would make a lower score than some one who knew very few of the French words, but knew English well enough to know the proper meanings of those few.

Variability in what is measured.—Another condition to be taken into consideration in educational as in many other measurements is that the thing being measured varies from time to time. In other words, even if the measuring instrument employed in a given case is perfectly reliable one cannot expect always to secure the same result from two or more applications thereof. It is a well known fact that this condition holds for measurements in physical science also. The length of a bar of metal, for example, varies according to its temperature, increasing when hot and decreasing when cold. Likewise an individual's height varies according to his physical condition, ordinarily being slightly greater in the morning than in the evening, and greater when he is fresh and vigorous than when he is tired. The same is true of mental abilities and traits. The pupil who is able to solve ten problems correctly in twenty-five minutes one day may do the same ten or a similar set in twenty-three minutes on another day, and in twenty-eight on another. Similarly a pupil who can give the events which occurred at twenty of twenty-five important historical dates at one time may at a later period have forgotten some one or more of the twenty previously given correctly, or without intervening study he may at the second period recall some of the five which he could not give on the first test. Therefore the unreliability of the scores obtained should not be charged entirely to the unreliability of the measuring instruments employed, but at least partially to differences

in the conditions under which the tests are given, especially in the mental, including the physical in so far as it affects the mental, tone and set of the pupils. The use of proper directions to pupils accompanied by the teacher's best efforts to secure uniform conditions from time to time will reduce this element of unreliability somewhat, but will not eliminate it.

Quantity and quality.—Quantity and quality are sometimes thought of as opposing elements which must be measured by radically different methods. This viewpoint, however, is seen to be unjustified when one looks at the actual situation. For most practical purposes quality is measured and expressed in quantitative terms. For example, we commonly speak of one pupil's translation of a passage in a foreign language as being better or of higher quality than that of another pupil. If we analyze the factors which contribute to better quality in the first case, we will find that they are at least largely quantitative. The first pupil knows the best English equivalents of more words than the second, he understands more thoroughly the inflectional endings used, the rules of syntax employed, and the other elements of which knowledge is necessary in translating. Even if no attempt is made to count items or points in determining scores, but merely "general merit" ratings given, these are usually in numerical or quantitative terms. This is usually the case when English composition, drawing, handwriting and other similar scales are employed. In only a few lines of educational measurement, of which perhaps the most prominent is the rating of character and personality, is any considerable use made of adjectives or descriptive phrases of quality or degree rather than numerical quantities.

The statement could also be made that most quantitative measures involve or imply quality. A pupil who can solve correctly twice as many algebraic problems as another pupil, both working the same length of time upon the same problems, is generally accredited with possessing a higher quality of ability in algebra.

The dimensions of pupils' performances.—Just as we refer to the dimensions of a physical object, so the same expression is frequently applied in connection with educational measurement. Ordinarily there are three such dimensions of pupils' performances or traits which should be measured. These are amount

done in a given time or rate of work, quality or accuracy of performance, and character or type of performance, frequently expressed as degree of difficulty. Thus, for example, if we wish to measure and describe a pupil's ability in algebra, we should take into account how rapidly he works, how accurate his work is, and the kind or difficulty of the examples or problems dealt with. Not all measuring instruments involve all three dimensions, however, many of them omitting the first. In many cases pupils are given enough time or practically enough to respond as well as they can to the exercises set before them, so that their scores are limited by the difficulty of the material and the correctness of their responses. For most tests, whether rate is an important element or not, a single score that involves in varying proportions the three dimensions is computed. In some cases two or even three separate scores are employed. Usually, however, if there is a uniform time limit and credit is given only for correct responses, a single score describes a pupil's performance well enough for most purposes. It can readily be seen, however, that such a score does not yield diagnostic information. For example, a pupil who attempted ten examples in algebra and got them all right, and another who attempted twenty and got only ten right would ordinarily both receive a score of ten. Similarly one who solved the ten easiest problems in the test and another who solved the ten most difficult would often receive the same score. Generally, however, the arrangement of the test and the directions given pupils are such that the discrepancy due to the latter cause is not very great. The failure to discriminate between those who work slowly but correctly, and those who work more rapidly but with less accuracy is, however, a very common fault of tests.

Types and varieties of educational measuring instruments.—So far the writer has spoken of the educational measurements movement more or less as if it were a unified whole, as though it were not subdivided into rather large phases. This, however, is not the case. Reference has already been made to the fact that tests have been developed for measuring performances in the school subjects, general intelligence, personality, and various other abilities and traits of pupils. Likewise there are rating devices of one sort or another, such as scales, score cards, and so

forth, for measuring many other matters with which education is concerned. The general movement might, therefore, be broken up into a number of parts on this basis. It is not, however, the writer's intention to do so. Instead he wishes to deal merely with that portion of the movement which concerns itself directly with the measurement of pupils and their characteristics, and to define certain general types or varieties of instruments used for this purpose.

Standardized tests.—In the first place, so-called objective tests may be divided into two groups, standardized tests and the new examination. Originally and in its narrowest sense "standardized" was applied to a test that had been widely enough given that the results therefrom indicated what might be expected of pupils of a given age, grade, or other homogeneous group. In general usage, however, the adjective "standardized" or "standard" also implies that the test in question has been carefully constructed according to certain general principles which will be dealt with in some detail later, and embodies exercises of such forms that pupils' responses are relatively, if not absolutely, objective. Furthermore, practically all tests which merit the name standardized are commercially available, that is, may be purchased from a publisher by anyone desiring to do so.[15]

The new examination.—The new examination or, better, the new-type test, is the name commonly given to tests or exercises, generally constructed by a teacher for her own use, that make use of the forms and scoring methods of standardized tests so as to possess relatively high objectivity, but have not gone through a process of careful trying out of material included, have not been given to large numbers of pupils, and are not generally available for use by others. They include true-false statements and yes-no questions, single-answer questions, multiple-answer exercises, matching exercises, completion statements, and other similar types.[16] No sharp distinction can be drawn between the two general varieties, since there are tests in all stages of de-

[15] Most of the standardized tests not commercially available at present have been so at some time, but are no longer on the market because other and better tests have since appeared.

[16] For descriptions of these and other varieties of new-type tests, see Chapter XX.

velopment from an ordinary new-type test constructed by a teacher for use with a single class up to a thoroughly standardized test. Both are thought of as opposed to traditional examinations, the primary difference being that they call for very brief pupil responses and are objective or nearly so.

Intelligence tests.—The two general classes of standardized tests most commonly employed in school work are general intelligence, or just intelligence, tests, and achievement tests. It has been suggested with considerable justification that the application of the term "intelligence" to most, if not all, tests commonly so designated is a misnomer and that a better expression would be "mental alertness tests," "tests of ability to do academic work," or something else. Such suggested terms have not been widely employed and do not appear likely to be in the near future. Regardless of the designation employed, however, one should bear in mind that all, or practically all, tests of this sort measure a combination in unknown proportions of inherited and acquired intelligence. Furthermore, they do not measure all phases of intelligence but merely certain samplings or manifestations thereof that are usually of more or less the same abilities as are required to do successful work, especially academic work, in school.

Achievement tests.—An achievement test is, of course, one which measures a pupil's achievement or performance along some line, ordinarily one of the school subjects. "Accomplishment test," "attainment test" and "subject-matter test" are frequently used in place of "achievement test." Sometimes the term "educational test," which has already been defined in a broader way, is employed in the same sense, but it is best not to do so.

General survey tests.—On another basis tests, especially achievement tests, may be thought of as being divided into three varieties: general survey tests, diagnostic tests, and prognostic tests. A "general survey," sometimes shortened to "general" and sometimes to "survey," test is one that yields merely a general or all-around measure of achievement in one or perhaps in several school subjects. Usually such a test includes one or a few exercises on each of the chief divisions or phases of the subject or subjects dealt with. For example, such a test in Latin

would probably include a few exercises dealing with each of the following phases, and perhaps others: vocabulary, declension, comparison, conjugation, grammatical rules, translation. Sometimes an intelligence test is included along with several achievement tests under the name of a general survey test or battery of tests.

Diagnostic tests.—A diagnostic test is, as its name implies, one which may be used to diagnose a pupil's capacity or performance. Therefore it must yield relatively detailed information concerning pupil capacity or achievement in one or more limited fields. The extent to which a test is valuable for diagnostic purposes depends largely on how narrowly limited the field or fields dealt with are, and how completely each is covered. For example, a Latin test such as that suggested in the last paragraph, which contains several subtests dealing with different phases of the subject, is to some extent diagnostic if the various subtests are long enough to yield separate scores of fair reliability. If this is carried further and the vocabulary subtest, for example, divided into parts which deal with nouns, adjectives, verbs, and so forth, or with groups of words divided on some other basis, the results yielded have still more diagnostic value. Tests that may be given within any reasonable length of time cannot be made completely diagnostic, except over very small portions of subject-matter, since doing so involves covering every item of knowledge or every elemental ability supposed to have been acquired. Thus such an outcome as knowing the names and dates of the presidents of the United States, or how to solve simple equations of the type $ax + b = c$, can be thoroughly tested for diagnostic purposes in a single test, but not knowledge of the lives and influences of the presidents or of how to solve all sorts of equations. For these latter, series of tests covering considerable time would be needed.

Prognostic tests.—A prognostic test differs from both of those just defined in that its function is not to measure pupils' achievement in school subjects, but rather to predict what their probable achievement or success in school, in a vocation or anywhere else, will be. Practically all achievement tests have some prognostic value, although they are not intended primarily for this purpose. There are, however, tests which have this as their primary,

and in some cases only, function. This is to some extent true of intelligence tests, but more so of a limited number prepared for use in connection with certain school subjects that do not chiefly if at all measure previous achievement therein. Such tests often consist of a series of short lessons in the subject, each followed by a test to see how well it has been learned.

Practice tests.—A fourth variety of tests in the school subjects is commonly known as practice tests, or sometimes as instructional tests. Such tests may be standardized, but usually are not. There are usually a large number of tests in a practice series, sometimes almost one for every day of the school year, and usually at least one or two for each week. Frequently, but not always, provision is made for pupils to correct their own work, so that the tests can be used for practice or study purposes without consuming much of the teacher's time. In many cases such series of tests differ little if any from the series of exercises or problems which are given in algebras, first-year foreign language books, and others, to be answered by the pupils.

"Tests" and "Scales."—So far in this volume the word "test" has been used in a broad sense to refer to all sorts of instruments used for purposes of measurement. The writer will to some extent continue to employ it in the same sense, especially in the expression "standardized test," but there is also a narrower meaning with which the reader should be familiar. In this sense, "test" as opposed to "scale" refers to a measuring instrument or the portion thereof intended to secure pupil performances that may be measured or rated. In other words, it is the questions, exercises or problems, with necessary directions and so forth, to which pupils respond in writing or otherwise. "Scale," on the other hand, refers to a measuring instrument or the portion thereof which is used in describing pupils' performances. In some cases a test and a scale are combined in a single instrument, in others they are separate. For example, a series of equations to be solved, such as $2x = 10$; $3x - 2 = 7$; $5x + 4 = 11x - 32$; $3x^2 - 11x - 77 = 27$; and so on, each more difficult than the one preceding it and assigned a value or weight on the basis of this difficulty, constitute a test in that they secure from the pupil his performance, that is, his responses to the exercises or attempts to give the correct answers. On the

other hand, they likewise constitute a scale, since a pupil's score is determined by the weights of the exercises which he solves correctly, or perhaps merely by that of the most difficult one he solves. In such a subject as handwriting the test and the scale are more distinct. The former consists of a situation presented to the pupil that secures from him a sample of his handwriting, presumably under standard or definitely determined conditions, whereas the scale consists of a series of specimens of handwriting ranging from very poor to very good, with which a pupil's performance is compared to determine its degree of merit or score. Although the word "scale" is in general use to include both of these types of devices for measuring pupils' performances, some writers have advocated that it be restricted to the second type of instrument, that is, to the one composed of samples or specimens arranged in order of merit with which pupils' work is compared.[17]

"Scaled tests" and "rate tests."—Most of those who advocate the restricted use of "scale" just stated employ the term "scaled test"[18] to refer to a test composed of exercises arranged in order of increasing difficulty. Although not yet generally adopted the use of this expression in this sense is here recommended. The term "rate test" or, less frequently, "speed test" is frequently employed in contrast to refer to one that has such a short time limit that few if any pupils have time enough to give all the correct responses they could if more time were allowed. The reason is that there is a general tendency for the time limits of unscaled tests to be short enough that the number of correct responses given by pupils is limited by the time allowed, whereas in the case of scaled tests the number of correct responses given is usually limited by the pupil's inability to respond correctly to the more difficult exercises. Also the terms "uniform test" and "irregular test" are used in much the same sense as "rate test." The former is employed because of the assumption, for scoring purposes at least, that if a test

[17] The expression "quality scale" is not infrequently used in this sense, that is, as synonymous with the restricted definition of scale suggested above.

[18] "Power test" and less frequently "difficulty test" are terms sometimes employed in the same sense as "scaled test."

is not scaled the exercises are of uniform or equal difficulty. "Irregular test" is employed by virtue of the fact that if exercises are not arranged in scalar order and are, as is usually the case, not of uniform difficulty, it is common to arrange them irregularly with regard to difficulty. Thus a typical irregular test in solving simple equations might contain the following series of exercises: $3x + 4 = 19$; $\frac{5x}{2} - 6 = x + 6$; $2x + 1 = 6x - 23$; $9x + 16 = \frac{14x}{3} + 29$; $5x = 2x + 21$; and so on.

Individual and group tests.—All tests, whether of intelligence, achievement or anything else, and regardless of their particular function, may be classed as either individual or group tests. That is to say, they are designed either to be given to a single individual at a time or to a group of individuals. Although they usually yield more valid and reliable scores than do group tests, comparatively few individual tests are widely employed. This is because of the large amount of time consumed in their use. The individual tests most commonly employed in school work are of intelligence, in which, however, they are not nearly as numerous as the group ones, and in such subjects as oral reading and public speaking in which the desired responses are oral. They are also used to some extent in the manual or mechanical subjects.

Cycle and spiral tests.—Among the varieties of tests from the standpoint of the arrangement of exercises or items, are cycle and spiral tests. A cycle test is one in which the same kinds of exercises or items recur in cycles, ordinarily in regular cycles. For example, a foreign language vocabulary test in which the first word is a noun, the second an adjective, the third a verb, the fourth a noun again, the fifth a second adjective, and so on in rotation, is a cycle test. Similarly in literature, if a test contains one question dealing with *Ivanhoe,* then one with *David Copperfield,* one with *Macbeth,* one with *The Crisis,* a second with *Ivanhoe,* and so on, it is also a cycle test.

A spiral test is similar to a cycle test except that it involves scalar arrangement also. In other words, it is a cycle test in which the items increase in difficulty. Thus a spiral test in algebra might have such exercises as the following:

1. $2x = 4$
2. $3x - 2 = 7$
3. $4y + 1 = 6y - 7$
4. $\frac{3x + 2}{5} = \frac{5x + 6}{9}$
5. $7x = 21$
6. $11y - 19 = 14$
7. $18x + 13 = 24x - 1$
8. $\frac{14y + 17}{9} = \frac{26y + 23}{15}$

It will be seen that No. 5 is similar to No. 1, but more difficult. The same is true of Nos. 6 and 2, 7 and 3, and 8 and 4. If the test were continued No. 9 would, of course, be similar to Nos. 5 and 1, but still more difficult. Occasionally a spiral test consists of groups of items or exercises arranged in spiral order. Thus there might be first a group similar to No. 1 in difficulty as well as form, then a group similar to No. 2, and so on. The chief advantage of the cycle or spiral arrangement is that it insures that an individual taking the test encounters at least some items of each kind, which might not be the case if all of one kind were grouped first, then all of a second kind, and so on, since with this arrangement he might leave unattempted all of those in one or more groups toward the end of the test. If the cycle is regular, it renders the test results more usable for diagnostic purposes, as the teacher can then more easily determine separate scores for different phases of the subject or types of exercises.

Verbal and non-verbal tests.—One general basis on which all tests may be divided into two groups is whether or not they are verbal. This term is used in two somewhat different senses. In one it refers to any test in which either the examiner or the persons being tested employ spoken or written language; in the other it refers only to those in which the subject has to read written directions or make use of language in his responses. Conversely, a non-verbal or non-language test in the strictest sense is one in which neither tester nor testee makes any use of written or spoken language, but in its broader sense includes tests in which the examiner gives oral directions but not those in which the subject must read directions or respond in words. Almost all subject-matter tests are verbal tests, practically the only exceptions being a very few in the case of such subjects as shop work, mechanical drawing, and so forth. Most intelligence tests also are verbal, but the proportion of these that is

non-verbal is considerably greater than in the case of achievement tests. Non-verbal tests are also sometimes used in connection with determining vocational capacity or ability.

Summary.—In this chapter an endeavor has been made to indicate the place of measurement in education. The purpose of the book is stated, also the purposes of educational tests. It is shown that more accurate measurements are needed largely because of the subjectivity of ordinary school marks. The causes of this subjectivity are indicated and briefly discussed. Following this the benefits derived from employing objective measurements are pointed out and eight of the most frequent objections to such measurements considered. The problem of securing performances which can be observed is discussed, also the variability of what is being measured, the place of quantity and quality in educational measurement, and the three dimensions of pupils' performances which should be measured. Finally the following varieties of educational measuring instruments are defined or explained: standardized tests; the new examination; intelligence and achievement tests; general survey, diagnostic, prognostic, and practice tests; tests as distinguished from scales, including quality scales, scaled tests, and rate tests; individual and group tests; cycle and spiral tests; and verbal and non-verbal tests.

References [19]

Carroll, R. P. "The Need and Advantages of Objective Measurements," *Fundamentals in the Technique of Educational Measurements.* Syracuse; New York: Author, University Station, 1928, Chapter I.

Gilliland, A. R. and Jordan, R. H. "Reasons for Educational Measurements," *Educational Measurements and the Classroom Teacher.* New York: The Century Co., 1924, Chapter I.

Greene, H. A. and Jorgensen, A. N. "Introduction" and "The Meaning of Educational Tests," *The Use and Interpretation of Educational Tests.* New York: Longmans, Green and Company, 1929, Chapters I and II.

Gregory, C. A. "Introduction" and "Efficiency through Measurements,"

[19] The lists of references given at the ends of the chapters consist of a relatively small number of selected references which the writer believes will be most helpful for further reading upon the topics dealt with by the various chapters. The references given in the footnotes within the chapters are not repeated in the lists at the ends unless they are helpful in connection with other points than those for which they were previously given.

Fundamentals of Educational Measurement. New York: D. Appleton and Company, 1922, Chapters I and II.

McCall, W. A. "Place of Measurement in Education," *How to Measure in Education.* New York: The Macmillan Company, 1922, Chapter I.

Monroe, W. S. "Nature and Process of Educational Measurements," *An Introduction to the Theory of Educational Measurements.* Boston: Houghton Mifflin Company, 1923, Chapter II.

Monroe, W. S., DeVoss, J. C., and Kelly, F. J. "Introduction," *Educational Tests and Measurements,* Revised and Enlarged Edition. Boston: Houghton Mifflin Company, 1924, Chapter I.

Odell, C. W. *Traditional Examinations and New-Type Tests.* New York: The Century Co., 1928, p. 1–40.

Ruch, G. M. "Points of View," "Objections to the Traditional Examination," and "Advantages and Limitations of Objective Examinations," *The Objective or New-Type Examination.* Chicago: Scott, Foresman and Company, 1929, Chapters I, III, and IV.

Symonds, P. M. "Why Measurement in High School?" and "Why Better Measurement in High School?" *Measurement in Secondary Education.* New York: The Macmillan Company, 1927, Chapters I and II.

Trabue, M. R. "The Nature of Measurements" and "Classifications of Mental Tests," *Measuring Results in Education.* New York: American Book Company, 1924, Chapters I and XIV.

CHAPTER II

THE HISTORY AND PRESENT STATUS OF THE EDUCATIONAL MEASUREMENT MOVEMENT[1]

Beginnings prior to 1900.—As was stated in Chapter I, educational measurement of some sort or other has existed practically ever since education began. Tests and examinations of various kinds were in use hundreds and even thousands of years ago among such peoples as the Chinese, the Greeks, and the Romans. For at least several centuries, if not longer, written and oral examinations similar in many respects to those commonly employed in our schools today have been in use. Needless to say, there have been changes in the methods of measurement from time to time, but what is generally considered the greatest departure from past practices along this line is a movement of very recent origin. As is true in the case of many movements, no definite date can be assigned as the exact beginning of what is frequently called the "educational measurement" or, more narrowly, the "standardized test" movement. Practically all that has been done along this line, however, has occurred within the last quarter of a century, although a few events prior to that time should be mentioned in connection with it.

For many years some teachers have, either occasionally or regularly, made use of some of the varieties of objective or near-objective exercises now employed in standardized tests. Moreover tests or examinations were at times given to fairly large numbers of pupils, sufficient perhaps to establish satisfactory norms.[2]

[1] A somewhat more detailed account will be found in "Research in Educational Measurement," Chapter IV of Monroe, W. S., et al. "Ten Years of Educational Research, 1918–27," *University of Illinois Bulletin*, Vol. 25, No. 51, Bureau of Educational Research Bulletin, No. 42. Urbana: University of Illinois, 1928. 367 p.

[2] "Norm" is the term used to refer to the statement of the actual achievement of a group of pupils which is homogeneous in some one respect. Usually the norm is expressed in terms of the median score. The median, commonly abbreviated Md., is the point which divides a number of scores into

Very little, if any, of such work was described in print or attracted much attention. Indeed, the only definitely reported work of this sort occurring more than a generation ago seems to be that of an English schoolmaster, Reverend George Fisher.[3] About 1864 he constructed a "scale book" which contained samples of typical questions and of various degrees of proficiency in answering the questions in several school subjects. The questions were intended to be models for the construction of future examinations similar in nature and difficulty. It is not apparent, however, that this work of Fisher's attracted any attention at the time, although it contained the germ of a number of principles later employed.

About this time Sir Francis Galton in England, and a few years later J. McKeen Cattell in America, began work along the lines of measuring individual differences and mental abilities which undoubtedly helped to prepare the way for the standardized-test movement, particularly the intelligence testing phase thereof. Their long-continued studies and experimental work, and their inspirational influence through writing and teaching, played a large part in causing later developments.

The first event which appears to have had any direct connection with the modern measurements movement was Dr. J. M. Rice's work which began in 1894 and continued through several years.[4] The best known part thereof had to do with a uniform spelling

two equal groups, or upon each side of which half of the scores lie. Thus if the norm for pupils studying first-year French is reported as being 30 points upon a particular test, it means that half of the pupils tested had scores at or above 30 and half at or below 30.

[3] Chadwick, E. B. "Statistics of Educational Results," *The Museum, A Quarterly Magazine of Educational Literature and Science*, 3: 479–84, January, 1864.

[4] Rice, J. M. "The Futility of the Spelling Grind," *Forum*, 23:163–72, 409–19; April–June, 1897.

———— "Educational Research: A Test in Arithmetic," *Forum*, 34: 281–97, October–December, 1902.

———— "Educational Research: Causes of Success and Failure in Arithmetic," *Forum*, 34: 437–52, January–March, 1903.

———— "Educational Research: Talent vs. Training in Teaching," *Forum*, 34: 588–607, April–June, 1903.

———— "Educational Research: The Results of a Test in Language," *Forum*, 35: 269–93, October–December, 1903.

test which he gave to pupils in a number of cities. In addition he constructed and administered tests in arithmetic and language. When he first reported his results, the attitude of the majority of educators was hostile and no immediate results of value followed. Fifteen years later, after the movement had got well under way, Rice was accorded considerable honor for his pioneer work.

The first book in this field.—The date, 1904, is sometimes given as marking the beginning of the modern educational measurements movement, because of the fact that in that year appeared the first book dealing directly with mental measurement. This volume, which was written by Thorndike,[5] dealt with statistical methods and fundamental principles of test construction. At once it began to be employed as a text and reference book, and to stimulate activity in this field. For ten years it remained practically the only one of its kind, but now is of little interest except historically.

The Binet-Simon Scale and its revisions.—Another noteworthy event occurred the next year. This was the publication of the first version of the now well-known Binet-Simon General Intelligence Scale.[6] This individual scale combined tests of a number of different kinds into a single scale, and provided for interpreting pupils' answers in terms of mental age.[7] Though it marked a noteworthy step, many imperfections were found, with the result that revisions appeared in 1908 and 1911.[8] This

———— "English, the Need of a New Basis in Education," *Forum*, 35: 440–57, January–March, 1904.

———— *Scientific Management in Education*. New York: Hinds, Noble and Eldredge, 1912. Chapters V to XI contain the same material as is given in the preceding six references.

[5] Thorndike, E. L. *An Introduction to the Theory of Mental and Social Measurements*. New York: Teachers College, Columbia University, 1904. 277 p. (Revised edition, 1913.)

[6] Binet, A. et Simon, T. "Methodes Nouvelles pour le Diagnostic du Niveau Intellectuel des Anormaux," *L'Année Psychologique*, 11:191–244, 1905.

[7] Mental age, abbreviated M.A., is a pupil's score on an intelligence test expressed in terms of age. To say that a pupil has a mental age of any given amount—for example, 14 years 8 months—means that his intelligence test score is the same as the average score made by pupils of that age.

[8] Binet, A., et Simon, T. "Le Développement de l'Intelligence chez les Enfants," *L'Année Psychologique*, 14:1–90, 1908.

scale, which will be described more fully later,[9] has been translated either literally or with considerable modification into many languages and is generally accepted as the best intelligence scale yet produced. It received little attention in this country until 1908, when Goddard[10] began to employ a fairly exact translation. In 1911 he published a revised form,[11] and the following year Kuhlmann issued another.[12] Although these revisions possess considerable merit, and a number of others have appeared since, the Stanford Revision by Terman and others is generally considered the best individual intelligence scale in the English language.[13] It first appeared in 1912,[14] but was not made generally available until 1916.[15]

Binet, A. "Nouvelles Recherches sur la Mesure du Niveau Intellectuel chez les Enfants d'École," *L'Année Psychologique*, 17:145–201, 1911.

[9] See p. 395.

[10] Goddard, H. H. "Four Hundred Feeble-Minded Children Classified by the Binet Method," *Pedagogical Seminary*, 17: 387–97, September, 1910.

———— "Two Thousand Children Measured by the Binet Measuring Scale of Intelligence," *Pedagogical Seminary*, 18: 232–59, June, 1911.

[11] Goddard, H. H. "A Revision of the Binet Scale," *Training School Bulletin*, 8: 56–62, June, 1911.

[12] Kuhlmann, F. "A Revision of the Binet-Simon System for Measuring the Intelligence of Children," *Journal of Psycho-Asthenics, Monograph Supplement*, Vol. 1, No. 1, September, 1912. The same author also published a later and better revision, for which see Kuhlmann, F. *A Handbook of Mental Tests*. Baltimore: Warwick and York, 1922. 208 p. Still more recently, in conjunction with Anderson, he has prepared another series of tests. See Kuhlmann, F. "A Median Mental Age Method of Weighting and Scaling Mental Tests," *Journal of Applied Psychology*, 11: 181–98, June, 1927.

[13] More recently, in 1922, Herring published an excellent revision which is generally considered of almost if not quite equal merit with the Stanford Revision. See Herring, J. P. *Herring Revision of the Binet-Simon Tests*. Yonkers: World Book Company, 1922. 56 p.

[14] Terman, L. M. and Childs, H. G. "A Tentative Revision and Extension of the Binet-Simon Measuring Scale of Intelligence," *Journal of Educational Psychology*, 3: 61–74, 133–43, 198–208, 277–89; February, March, April, May, 1912.

[15] Terman, L. M. *The Measurement of Intelligence*. Boston: Houghton Mifflin Company, 1916. 362 p.

Terman, L. M. et al. *The Stanford Revision and Extension of the Binet-Simon Scale for Measuring Intelligence*. Baltimore: Warwick and York, 1917. 179 p.

The first standardized achievement tests and scales.—Although the first intelligence scale appeared in 1905, it was not until 1908 that anyone followed up Rice's work by publishing a standardized test or scale in any school subject. In this year Stone, a student of Thorndike's, issued his arithmetic reasoning test.[16] This is generally considered the first standardized subject-matter or achievement test. For the next few years standardized tests and scales appeared at the rate of about one a year, practically all of them being by Thorndike and his students. As was probably to be expected, these early tests and scales were for use in subjects entirely or primarily taught in elementary rather than high school. The following list includes those which had appeared by 1913, also one noteworthy but somewhat later one: Courtis Arithmetic Tests, Series A, (1909) [17]; Thorndike Scale for Handwriting of Children (1909) [18]; Hillegas Scale for the Measurement of Quality in English Composition by Young People (1912) [19]; Buckingham Spelling Scale (1913) [20]; Thorndike Scale for General Merit of Children's Drawings (1913) [21]; Ayres Scale for Measuring the Quality of Handwriting of School Chil-

[16] Stone, C. W. "Arithmetical Abilities and Some Factors Determining Them," *Teachers College, Columbia University, Contributions to Education*, No. 19. New York: Bureau of Publications, Teachers College, Columbia University, 1908. 101 p.

[17] Courtis, S. A. *Manual of Instructions for Giving and Scoring the Courtis Standard Tests in the Three R's.* Detroit, Michigan: Department of Co-operative Research, 1910.

Courtis' Standard Research Tests in Arithmetic, Series B, which cover the four fundamentals and probably have received the widest use of any standardized tests, were not prepared until four or five years later than his Series A.

[18] Thorndike, E. L. "Handwriting," *Teachers College Record*, 11:1–93, March, 1910.

[19] Hillegas, M. B. "A Scale for the Measurement of Quality in English Composition by Young People," *Teachers College Record*, 13: 331–84, September, 1912.

[20] Buckingham, B. R. "Spelling Ability: Its Measurement and Distribution," *Teachers College, Columbia University, Contributions to Education*, No. 59. New York: Bureau of Publications, Teachers College, Columbia University, 1913. 116 p.

[21] Thorndike, E. L. "A Scale for Measuring Achievement in Drawing," *Teachers College Record*, 14: 345–82, November, 1913.

dren (1912) [22]; Ayres Measuring Scale for Ability in Spelling (1915).[23]

It should be noted that two types of measuring instruments are represented among those named above. The Stone and Courtis Arithmetic Tests belong to one and the rest to the other. The former are tests in the narrowest sense of the word, that is, they consist of exercises or examples to which pupils respond by giving, or attempting to give, the answers. The other instruments mentioned are scales, that is, they consist of specimens of handwriting, composition, and so forth, arranged in order of merit; or, in the case of spelling, of words in order of difficulty, and are used to rate pupils' performances, though not directly to secure them. From about 1913 on, the number of tests increased so that it is not practicable or desirable to attempt to mention all of them. It was not long until some were available in all of the more fundamental elementary-school subjects, and a beginning made in high-school subjects.

Contributing causes to the growth of the movement.—In connection with this rapid increase in the number and use of standardized tests, several events and movements which served more or less as contributing causes thereto should be mentioned. One of these which deserves prominent mention was the considerably increased interest in school marks during 1910 and the few years immediately following. Although there had been discussions and studies of marks before this time, this period was marked by the appearance of reports of a number of studies [24] showing that school and examination marks as ordinarily given were decidedly subjective and, therefore, unreliable. The findings reported, which emphasized, probably even over-emphasized, the inaccuracy of ordinary school marks, soon attracted considerable attention and resulted in arousing much interest in

[22] Ayres, L. P. "Scale for Measuring the Quality of Handwriting of School Children," *Russell Sage Foundation Bulletin* E-113. New York City: Russell Sage Foundation, 1912. 16 p.

[23] Ayres, L. P. "A Measuring Scale for Ability in Spelling," *Russell Sage Foundation Bulletin* E-139. New York City: Russell Sage Foundation, 1915. 56 p.

[24] The best-known of these studies have already been listed on p. 6.

standardized tests as a probable means of securing more satisfactory ratings of pupils.

Another directly influential movement was the development of school surveys. This term seems to have been used first in connection with a study of the Pittsburgh schools in 1907, at which time there were, of course, no standardized achievement tests. The first survey to employ such tests was that of New York City in 1911–12.[25] Those making school surveys at once began to exert a strong demand for satisfactory tests of achievement and later of intelligence, and this demand appears to have been one of the strongest stimulants to test production. Few surveys of any note have been made within the last fifteen years in which standardized tests have not been employed.

A third factor that promoted the construction of tests was that several important periodicals began to devote considerable attention to their development. The *Teachers College Record*, the *Journal of Educational Psychology*, *Educational Administration and Supervision*, and *School and Society* may be mentioned as noteworthy in this respect, though many others also contained articles dealing with tests from time to time. More recently the *Journal of Educational Research*, which was not founded until 1920, has taken a very prominent position in this respect.

The development of the educational measurement movement was also greatly stimulated by addresses and test demonstrations at teachers' meetings of various sorts. Probably the Indiana University Conference on Educational Measurements is the most outstanding example of such meetings. Such a conference has been held every year, beginning in 1914, and the series has included among the speakers practically all the outstanding workers in the field. Many other universities have conducted similar meetings, many state, county and city teachers' organizations have furnished occasions for addresses, and many courses in teacher training have done much to develop interest and understanding.

When the first educational research bureaus were organized about 1912, their attention was at first almost entirely centered on the construction, use, and popularization of tests. Practically

[25] "Final Report of Committee on School Inquiry, Board of Estimate and Apportionment." New York City: The Committee, 1911–1913. 3 vols.

all of the numerous state, city, university and other such bureaus have continued to devote attention to this field, although in general the scope of their activities has been very much broadened.

The development of group intelligence scales.—Although, as has been stated above, the first individual intelligence scale appeared in 1905, and the first standardized achievement test in 1908, it was not until the time of our entrance into the World War that a group intelligence scale became available. Otis, working under Terman, is generally given credit for constructing the first scale of this sort,[26] although Pintner,[27] as well as others, had made some use of several group tests that tended to measure general intelligence. These tests, however, were not combined into a single unified scale. Because of the war Otis' scale did not appear as such until 1918, although it formed in large part the basis of the well-known Army Alpha Scale,[28] which was the one most commonly used in testing recruits. To supplement this verbal scale, the non-verbal Army Beta Tests [29] and also various others were constructed. The testing work in the army, which was under the direction of prominent psychologists, constituted the most extensive mental testing program the world has yet seen, and served as a very great stimulus to the carrying on of such activity in school, industry, and elsewhere. In the army testing the chief purpose was to determine the general intelligence of recruits, but also much was done by way of attempting to measure vocational abilities. Almost at once after the various army tests and the Otis scales became known, numerous group intelligence scales appeared, many attempts to measure educational and vocational aptitudes and abilities were made, and a decidedly increased number of achievement tests were published. Indeed, the stimulation from the army work was probably greater than desirable, and led to considerable non-critical ac-

[26] Otis, A. S. "An Absolute Point Scale for the Group Measurement of Intelligence," *Journal of Educational Psychology*, 9: 239–61, 333–48; May, June, 1918.

[27] Pintner, Rudolf. "A Mental Survey of the School Population of a Village," *School and Society*, 5: 597–600, May 19, 1917.

[28] Yerkes, R. M. (Editor). *Psychological Examining in the United States Army, Memoirs of the National Academy of Sciences*, Vol. 15. Washington: Government Printing Office, 1921. 890 p.

[29] *Op. cit.*

tivity, both in the construction of tests and in their use.

General survey tests.—At practically the same time that marked the beginning of group intelligence tests, general survey tests also began to be constructed and employed. The first test of this sort seems to have been that of Pintner.[30] It consisted of portions of eight already standardized achievement tests. In connection with its use and that of an intelligence test, Pintner made the first suggestion [31] of a definite measure for comparing pupils' achievements with their intelligence [32] or, in other words, for comparing how well they achieve in school with their capacity to achieve.

A little later, in 1920, Buckingham and Monroe published the Illinois Examination.[33] Instead of covering a relatively large number of subjects, this dealt merely with reading and arithmetic in addition to intelligence, thus tacitly arguing that a fairly satisfactory measure of pupils' general achievement can be obtained by testing the two most important subjects. It was in connection with this battery of tests that the now well known terms "achievement age" [34] and "achievement quotient" [35] were first employed.

[30] Pintner, Rudolf and Fitzgerald, Florence. "An Educational Survey Test," *Journal of Educational Psychology*, 11: 207–23, April, 1920.

[31] Pintner had suggested the desirability of comparing achievement with ability and had even made such comparisons two or three years earlier, but had not provided a unified battery of achievement tests nor suggested any very satisfactory means of making the comparisons.

[32] Pintner, Rudolf and Marshall, Helen. "A Combined Mental-Educational Survey," *Journal of Educational Psychology*, 12: 32–43, 82–91; January, February, 1921.

[33] Monroe, W. S. and Buckingham, B. R. *Illinois Examination. Teacher's Handbook*. Urbana: Bureau of Educational Research, University of Illinois, July, 1920. 32 p.

Monroe, W. S. "The Illinois Examination," *University of Illinois Bulletin*, Vol. 19, No. 9, Bureau of Educational Research Bulletin No. 6. Urbana: University of Illinois, 1921. 70 p.

[34] The expression "achievement age" (A.A.) is used to refer to a pupil's score on an achievement or subject matter test expressed in terms of age. An achievement age of a given amount—for example, 12 years 4 months—means that the pupil who earns this score exhibits ability equal to the average of pupils of the same chronological or mental age. The terms "accomplishment age," suggested by Franzen at about the same time, and "attainment age" are entirely synonymous with achievement age.

[35] The "achievement quotient" (A.Q.) is obtained by dividing the achieve-

Only a few months later the Presseys published their Scales of Attainment Nos. 1, 2, and 3,[36] each of which covered three subjects for the second, eighth, and third grades, respectively.

A year or so later, in 1922, appeared the outstanding general survey test to date. This is the Stanford Achievement Test,[37] which consists of a Primary Examination for Grades II and III covering reading, arithmetic and spelling, and an Advanced Examination for Grades IV to VIII, covering these and four or five other subjects. This test is outstanding because of its unusually careful construction and validation and its high reliability.

A number of other batteries of tests have appeared, but from a historical standpoint they are of little interest, and from that of worth so inferior to the Stanford Achievement Test as not to deserve mention. One exception among those intended for the elementary school is the Public School Achievement Tests by Orleans, of which Battery A covers reading, arithmetic, language, and spelling; B, grammar, history, and geography; and C, nature study and health.[38]

Within the past few years there have also appeared several batteries of tests designed for high-school use or in connection

ment age by the mental age; that is, $A.Q. = \frac{A.A.}{M.A.}$. Thus it compares a pupil's performance in a school subject with his capacity. The terms "accomplishment quotient," suggested by Franzen, and "attainment quotient" are also employed synonymously. Franzen later advocated using the term "accomplishment ratio" instead of "accomplishment quotient" and employing the latter in a different sense.

[36] Pressey, L. C. "Scale of Attainment No. 1.—An Examination of Achievement in the Second Grade," *Journal of Educational Research*, 2: 572–81, September, 1920.

Pressey, S. L. "Scale of Attainment No. 2.—An Examination for Measurement in History, Arithmetic, and English in the Eighth Grade," *Journal of Educational Research*, 3: 359–69, May, 1921.

Pressey, L. C. "Scale of Attainment No. 3.—For Measuring 'Essential Achievement' in the Third Grade," *Journal of Educational Research*, 4: 404–12, December, 1921.

[37] Kelley, T. L., Ruch, G. M., and Terman, L. M. *Stanford Achievement Test*. Yonkers, New York: World Book Company, 1922. (Revised, 1929.)

[38] Orleans, J. S. "Public School Achievement Tests, Batteries A, B, and C." Bloomington, Illinois: Public School Publishing Company, 1928.

with college entrance. The Iowa Placement Examinations [39] constitute the most complete series of this sort. They consist of six so-called training tests which measure achievement in as many of the most commonly taught high-school subjects, and five aptitude tests intended to predict ability to carry the same subjects.

Development of high-school and college tests.—It has already been stated that the development of high-school tests lagged behind that of elementary-school tests. Since about the time of the war, however, such tests have appeared in large numbers, until it is now true of high-school as well as elementary-school subjects that there are some tests in practically every one, and a rather large number in many. The development in the secondary field is not yet equal to that in the elementary field, however, and probably never will be. It was not until 1927 that the first two books devoted primarily to testing in the secondary field appeared, although most of the numerous books issued previous to this time which devoted most attention to the elementary field gave one or a few chapters to high-school subjects. The two books referred to, by Ruch and Stoddard [40] and Symonds,[41] respectively, are both of high merit and fairly critical in their attitude. That of Symonds is much the more inclusive of the two, devoting a number of chapters to various uses of tests and related questions with which Ruch and Stoddard deal briefly or not at all.

A beginning has also been made in preparing tests for use in institutions of higher learning. The tests now available not only cover subjects likewise taught in high school, such as algebra, foreign language, geometry, history, physics, and so forth, but also subjects such as education, psychology and law, which are rarely if ever offered elsewhere than in colleges and universities. Intelligence tests suitable for college and university students and adults have likewise been prepared, as well as for those in high and elementary school and even for those of pre-

[39] Stoddard, G. D. "Iowa Placement Examinations," *University of Iowa Studies in Education*, Vol. 3, No. 2. Iowa City: University of Iowa, 1925. 103 p.

[40] Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction*. Yonkers, New York: World Book Company, 1927. 381 p.

[41] Symonds, P. M. *Measurement in Secondary Education*. New York: The Macmillan Company, 1927. 588 p.

school age. Among the best books dealing with measurement above the secondary school are Wood's *Measurement in Higher Education*[42] and MacPhail's *The Intelligence of College Students*.[43] The former gives an account of the use of the Thorndike Intelligence Examination and new-type tests at Columbia University, whereas the latter gives a comprehensive summary of the use of intelligence tests in a large number of institutions.

The new examination.—Although there is a close connection between what is called the new examination and standardized tests, the latter had been in existence for more than a decade before the formal beginning of the former. At the beginning of 1920 appeared an article by McCall,[44] which seems to have been the first published discussion along this line. Considerable interest was aroused at once, and it was not long until many persons were employing, experimenting with, and advocating this type of test. During the years since McCall's article was published, hundreds of articles, scores of chapters or sections of books, and a few whole volumes have been devoted to the new examination. Among the latter, at least six appear to deserve mention. Ruch's *The Improvement of the Written Examination*,[45] which appeared in 1924, reported several studies of new-type tests and gave considerable helpful discussion of their construction, use, and so forth. Two years later Russell's *Classroom Tests*[46] came from the press. It is a larger volume, only in part devoted to new-type tests, however. After another interval of two years, a considerably more complete volume by Odell,[47] dealing with both traditional or essay examinations and new-type tests, was published. The same year appeared Orleans and Sealy's *Objective*

[42] Wood, B. D. *Measurement in Higher Education*. Yonkers, New York: World Book Company, 1923. 337 p.

[43] MacPhail, A. H. *The Intelligence of College Students*. Baltimore: Warwick and York, 1924. 176 p.

[44] McCall, W. A. "A New Kind of School Examination," *Journal of Educational Research*, 1: 33–46, January, 1920.

[45] Ruch, G. M. *The Improvement of the Written Examination*. New York: Scott, Foresman and Company, 1924. 193 p.

[46] Russell, Charles. *Classroom Tests*. Boston: Ginn and Company, 1926. 346 p.

[47] Odell, C. W. *Traditional Examinations and New-Type Tests*. New York: The Century Co. 1928. 469 p.

Tests,[48] which contains a detailed and helpful account of what was done in one school district to develop an elaborate set of new-type examinations. The most recent books in this field are two by Ruch[49] and Ruch and Rice,[50] respectively. The first is the outstanding treatise on the new examination, presenting a great many experimental data concerning various significant phases in addition to many helpful suggestions on its construction and use. The second presents the best of several hundred new-type tests actually made by teachers and submitted in a nationwide competition.

Noteworthy books in the general field.—Reference has already been made to Thorndike's pioneer publication in this field. For more than ten years it stood alone,[51] but after that time a number of other books began to appear. The first book dealing with achievement tests was Starch's *Educational Measurements,*[52] dated 1916, which was chiefly devoted to reproducing and describing the then existing tests and scales. The same year there also appeared the *Fifteenth Yearbook of the National Society for the Study of Education,*[53] of which Part I dealt with achievement and intelligence testing, physical measurement, and the rating of school buildings.[54] The first book that may be said

[48] Orleans, J. S. and Sealy, G. A. *Objective Tests.* Yonkers, New York: World Book Company, 1928. 373 p.

[49] Ruch, G. M. *The Objective or New-Type Examination.* Chicago: Scott, Foresman and Company, 1929. 478 p.

[50] Ruch, G. M. and Rice, G. A. *Specimen Objective Examinations.* Chicago: Scott, Foresman and Company, 1929. 324 p.

[51] Whipple's *Manual of Mental and Physical Tests,* published in 1910, was concerned with psychological tests as distinguished from educational tests. See p. 4.

Whipple, G. M. *Manual of Mental and Physical Tests.* Baltimore: Warwick and York, 1910. 534 p. (Revised edition, 1914. Part I, 365 p. Part II, 336 p.)

[52] Starch, Daniel. *Educational Measurements.* New York: The Macmillan Company, 1916. 202 p.

[53] Strayer, G. D., et al. "Standards and Tests for the Measurement of the Efficiency of Schools and School Systems," *Fifteenth Yearbook of the National Society for the Study of Education,* Part I. Chicago: University of Chicago Press, 1916. 172 p.

[54] Terman's *The Measurement of Intelligence,* previously referred to on page 33, was also published in 1916. This, however, is a narrowly specialized book, since it deals with the administration of the Stanford Revision of the Binet-Simon Scale, and not with testing in general.

to have been at all a comprehensive treatise on the use of achievement tests was Monroe, DeVoss, and Kelly's *Educational Tests and Measurements,*[55] which came from the press in 1917. This probably had more influence than any other single volume in popularizing the use of tests and suggesting the proper methods of employing them. In the same year also appeared Rugg's *Statistical Methods Applied to Education,*[56] which was the first book to furnish workers with a fairly adequate treatment of the elementary statistical methods needed in connection with testing work.

What is sometimes thought of as the preliminary and introductory period of the educational measurement movement is well described and summed up in Part II of the *Seventeenth Yearbook of the National Society for the Study of Education,*[57] dated 1918. This volume, prepared by a committee of the National Association of Directors of Educational Research,[58] consists of thirteen chapters, each written by a leader in the field. It treats of the history, status, and purpose of measuring achievement, describes practically all of the then existing tests and scales, discusses the work of educational research bureaus, points out the practical uses of measurement, explains the elementary statistical methods needed, and closes with a very complete bibliography. Although this publication as a whole is outstanding, one sentence which seems to have first appeared in print therein is perhaps even better known than the whole volume. At the beginning of Chapter II may be found Thorndike's now well known dictum "Whatever exists at all, exists in some amount," a statement that has been the center of much controversy, but has been accepted by many workers in the field. Since about the close of the World War the number of books in the field of testing has been

[55] Monroe, W. S., DeVoss, J. C., and Kelly, F. J. *Educational Tests and Measurements.* Boston: Houghton Mifflin Company, 1917. 309 p. (Revised and enlarged edition, 1924. 521 p.)

[56] Rugg, H. O. *Statistical Methods Applied to Education.* Boston: Houghton Mifflin Company, 1917. 410 p.

[57] Courtis, S. A., et al. "The Measurement of Educational Products," *Seventeenth Yearbook of the National Society for the Study of Education,* Part II. Bloomington, Illinois: Public School Publishing Company, 1918. 192 p.

[58] This association has since changed its name to the American Educational Research Association.

so great, and likewise the contributions made by many of them so slight, that it is not desirable to mention all. A few, however should be referred to briefly. McCall's *How to Measure in Education*[59] and Monroe's *Introduction to the Theory of Educational Measurements*[60] were the first relatively advanced texts to appear. Both devote considerable attention to test construction and other topics not highly essential for the ordinary classroom teacher. Kelley's *Interpretation of Educational Measurements*[61] contains highly critical discussions of a few important problems and questions. It also has ratings of almost all the existing standardized tests of any merit by a number of experts in the field.

Three of the most recent books dealing with the general field of educational measurements are Smith and Wright's *Tests and Measurements,*[62] the revised edition of Wilson and Hoke's *How to Measure,*[63] and Greene and Jorgensen's *The Use and Interpretation of Educational Tests.*[64] The first two describe a large number of standardized tests, to some extent critically, and contain rather limited discussions of the use of tests for various school purposes and of related questions. The volume by Greene and Jorgensen is perhaps the best existing introductory discussion of educational testing, but does not contain descriptions of specific tests.

The *Twenty-First Yearbook of the National Society for the Study of Education*[65] was the first noteworthy publication devoted to intelligence tests in general. It dealt with the nature,

[59] McCall, W. A. *How to Measure in Education.* New York: The Macmillan Company, 1922. 416 p.

[60] Monroe, W. S. *An Introduction to the Theory of Educational Measurements.* Boston: Houghton Mifflin Company, 1923. 364 p.

[61] Kelley, T. L. *Interpretation of Educational Measurements.* Yonkers, New York: World Book Company, 1927. 363 p.

[62] Smith, H. L. and Wright, W. W. *Tests and Measurements.* New York: Silver, Burdett and Company, 1928. 549 p.

[63] Wilson, G. M. and Hoke, K. J. *How to Measure,* Revised and Enlarged. New York: The Macmillan Company, 1928. 597 p.

[64] Greene, H. A. and Jorgensen, A. N. *The Use and Interpretation of Educational Tests.* New York: Longmans, Green and Company, 1929. 389 p.

[65] Colvin, S. S., et al. "Intelligence Tests and Their Use," *Twenty-First Yearbook of the National Society for the Study of Education.* Bloomington, Illinois: Public School Publishing Company, 1922. 289 p.

history, and general principles of intelligence testing, as well as the practical use of tests and their results. Freeman's *Mental Tests* [66] is the outstanding treatise on the general field of mental [67] testing in a broad sense. It traces the history of the movement, describes all important available tests, and discusses critically the interpretation and use of test results. Thorndike's *Measurement of Intelligence* [68] and Dearborn's *Intelligence Tests, Their Significance for School and Society* [69] are also of high merit, but much more limited in scope. The first deals with certain important problems in the field of intelligence testing, chiefly reporting the experimental work of Thorndike and his associates thereon, whereas the second is essentially a summary and evaluation of the outstanding work having to do with the interpretation and use of intelligence tests.

A publication which should be mentioned, although it is not a book of the ordinary sort, is the *Bibliography of Educational Measurements* [70] compiled by Smith and Wright of the Bureau of Coöperative Research of Indiana University. This contains the best available list of achievement tests, with brief descriptions of each, and other useful information.

Before closing this list of important books in the educational measurement field, reference should be made to several in the closely related field of statistics. After Rugg's previously mentioned volume [71] in 1917 there was little published in book form until 1925, when his second work along this line,[72] Thur-

[66] Freeman, F. N. *Mental Tests.* Boston: Houghton Mifflin Company, 1926. 503 p.

[67] Although the term "mental test" logically refers to any test dealing with the functioning of the mind as distinguished from that of the body, yet it is commonly employed as synonymous, or almost so, with intelligence test. Sometimes it is employed, as by Freeman, to include tests of personality as well.

[68] Thorndike, E. L., et al. *The Measurement of Intelligence.* New York: Bureau of Publications, Teachers College, Columbia University, 1927. 616 p.

[69] Dearborn, W. F. *Intelligence Tests, Their Significance for School and Society.* Boston: Houghton Mifflin Company, 1928. 336 p.

[70] *Bibliography of Educational Measurements.* Bloomington, Indiana: Bureau of Coöperative Research, Indiana University, 1923. 120 p. (First revision, 1925, 148 p.; second revision, 1928, 251 p.)

[71] See p. 43.

[72] Rugg, H. O. *A Primer of Graphics and Statistics for Teachers.* Boston: Houghton Mifflin Company, 1925. 142 p.

stone's,[73] Otis's,[74] and Odell's [75] appeared. Garrett's book [76] came off the press in 1926 and Holzinger's,[77] which is in many ways the best of those mentioned, in 1927.

Educational measurement in other fields.—Not only have standardized measuring instruments been developed in the fields of achievement and intelligence, but also in many others. In 1916 the first of the Strayer-Engelhardt series of score cards for rating school buildings [78] appeared. There are now available score cards by these or other authors for city elementary schools, high schools, rural schools, school administration buildings, school equipment, record and report systems, and so on. Their use has become a regular procedure in connection with school building programs and surveys.

During the last ten or twelve years the measurement of personality, character, emotions, attitudes, and so forth, has also received an ever increasing amount of attention. Much of the activity along this line has had to do with the construction and use of scales, but also a number of tests which may be administered to individuals have been constructed. Among the best known work in this field is that of Downey with her Will-Temperament Tests,[79] that of Voelker [80] in attempting to measure character and

[73] Thurstone, L. L. *The Fundamentals of Statistics.* New York: The Macmillan Company, 1925. 237 p.

[74] Otis, A. S. *Statistical Method in Educational Measurement.* Yonkers, New York: World Book Company, 1925. 337 p.

[75] Odell, C. W. *Educational Statistics.* New York: The Century Co., 1925. 334 p.

[76] Garrett, H. E. *Statistics in Psychology and Education.* New York: Longmans, Green and Company, 1926. 317 p.

[77] Holzinger, K. J. *Statistical Methods for Students in Education.* Boston: Ginn and Company, 1928. 372 p.

[78] Strayer, G. D. "Score Card for City School Buildings," *Fifteenth Yearbook of the National Society for the Study of Education,* Part I. Bloomington, Illinois: Public School Publishing Company, 1916, p. 41–51.

[79] Downey, J. E. "The Will-Profile, A Tentative Scale for Measurement of the Volitional Pattern," *Bulletin of the University of Wyoming,* No. 3. Laramie, Wyoming: Department of Psychology, University of Wyoming, 1919. 40 p.

[80] Voelker, P. F. "The Function of Ideals and Attitudes in Social Education," *Teachers College, Columbia University, Contributions to Education,* No. 112. New York: Bureau of Publications, Columbia University, 1921. 126 p.

finally Hartshorne and May's[81] along the same line. Numerous rating scales of school habits and attitudes such as neatness, attention, interest, initiative, perseverance, promptness, leadership, and so on almost ad infinitum have been devised, as well as similar instruments for various characteristics not particularly connected with school situations.

Another type of measurement in which there has been much activity in the past few years has been that of predicting vocational or occupational aptitudes. Some of this work has been in connection with guidance programs in school, but most of it has been concerned with the selection, placement, and promotion of employees by industrial, commercial, and other organizations. There are at present numbers of tests available for engineering, mechanical, and clerical aptitude as well as for actual ability in clerical and stenographic work. There are also, though in smaller numbers, tests available for bricklayers, chauffeurs, farmers, firemen, janitors, painters, policemen, storekeepers, telephone operators, and workers in many other occupations. In addition to tests designed particularly for such purposes, considerable use has been made of tests in some of the school subjects and still more of intelligence tests for the same purposes. The one outstanding volume in this field is undoubtedly Hull's *Aptitude Testing*.[82] This discusses in decidedly adequate fashion both the principles and methods of aptitude testing, not only with regard to vocations, but in a much wider field.

Finally, in addition to achievement and intelligence tests, score cards for school buildings, tests and scales for measuring personality, character, and so forth, and vocational tests, many other measuring instruments which have some connection with education have been devised. There have been published, for example, tests of creative ability, dramatic judgment, scientific thinking, study habits, and religious education, also score cards

[81] Hartshorne, Hugh, May, M. A., et al. "Testing the Knowledge of Right and Wrong," *Religious Education Association Monograph* No. 1. Chicago: Religious Education Association, July, 1927. 72 p. Also in *Religious Education*, February, April, August, October, December, 1926; May, 1927.

Hartshorne, Hugh and May, M. A. *Studies in Deceit. Studies in the Nature of Character*, I. New York: The Macmillan Company, 1928. 414 p.

[82] Hull, C. L. *Aptitude Testing*. Yonkers, New York: World Book Company, 1928. 535 p.

for rating textbooks, school records, and budget practices, community and home conditions, and so forth.

Two recent tendencies.—Any account of the development and present status of the educational measurements movement would not be complete without mentioning two somewhat recent tendencies. The first is that the use of standardized tests is no longer commonly thought of as an experiment or as something definitely apart from regular instructional and other activities. Instead it has come to be an integral part of the work of teachers in many systems, both large and small. Undoubtedly the most notable example of a large city system in which this is true is Detroit, where Courtis developed an extensive program, but in hundreds of others it has a prominent place. Some large cities, such as Philadelphia, Detroit, and Denver, construct many of their own standardized tests instead of purchasing them.

Not only do many city school systems carry on very complete testing programs, but state-wide, and occasionally even nation-wide, programs of this nature are carried on. It is nothing unusual for from fifty to one hundred thousand pupils or even more to be tested in a single subject at practically the same time. Indeed, on one occasion at least, more than half a million pupils from every state in the union except one, have participated in a single program.[88]

Another marked and very desirable tendency that was to be expected with time is the increased growth of the critical attitude. Tests are no longer generally given merely because it seems the thing to do or because so doing is supposed in some inexplicable way to improve the efficiency of instruction or the achievements of pupils. Instead those responsible for the administration of tests have definite purposes in mind and select their tests carefully in view of these purposes. Diagnostic tests, prognostic tests, practice tests, and others having specific functions are employed in ever increasing numbers, whereas tests which yield only general measures are taking a somewhat subordinate place. Those who are constructing tests give much more time and attention to the selection, formulation, and arrangement of items, the deter-

[88] *Report of the Fourth Annual Nation-Wide Testing Survey, Project No. 1, Intelligence Testing.* Bloomington, Illinois: Public School Publishing Company, 1927–28. 32 p.

mination of validity and reliability, and other important steps than was formerly true. More refined statistical techniques for dealing with test data are being employed and more emphasis placed upon interpretation.

The number of tests available and used at present.—In concluding this account of the history of the movement it seemed appropriate to give some idea of its extent at present. The writer endeavors to keep as nearly complete as possible a file of standardized and near-standardized tests. In this file at present he has about fifteen hundred different tests or series of tests.[84] Almost two hundred of these are general intelligence tests. The subjects of arithmetic, language, and reading have nearly one hundred each. Fields in which there are about fifty tests and scales each are character and personality, history, and teacher rating. Each of the following has approximately twenty-five or more: algebra, general survey tests, English composition, geography, geometry, home economics, Latin, English literature, mechanical ability, physics, stenography, religious education, vocabulary, and handwriting. Not all of the tests included in these figures are actually available for use. Some are no longer published because they have been superseded by other and better tests; others have never passed the experimental stage, and probably never will; others, though available, possess such slight merit that they are receiving, and deserve to receive, practically no use. About half of the measuring instruments included, however, are both commercially available at present and of high enough merit that there are situations in connection with school work in which they may profitably be employed.

In order to get information as to how widespread is the use of tests, the writer attempted to determine how many copies were sold in the United States during one year. All of the larger publishers and most of the smaller ones furnished information concerning their sales. From the figures given, it appears that not less than thirty million copies of standardized tests and scales, and

[84] A series of similar tests in the same subject by the same author is counted as only one. For example, if an author has prepared a series of reading tests of which one is for the lower, one for the intermediate, and one for the upper grades, or a series of intelligence tests including one for young children, one for adolescents, and one for adults, it has been counted as only one in each case.

perhaps close to forty million, were sold within one recent year. About one-fourth of the total number were intelligence tests, and three-fourths achievement tests. Included in the total were several tests for each of which the sales ran from five hundred thousand to a million or even more. In connection with these figures, it should be noted that many of the copies sold were of scales for use in such subjects as handwriting, drawing, and English composition, in which only the teacher needs a copy of the scale and this may be used for rating hundreds and even thousands of pupils' performances. Thus the total number of standardized scores or ratings given pupils' responses is undoubtedly much greater than the actual number of copies sold. Furthermore, the figures given are now two or three years old, and there seems no doubt that the number sold at present is considerably greater than at the time. Probably figures for the present would run from forty to fifty million.

Summary.—Although educational measurement has existed since time immemorial,the modern measurements movement has grown up within a generation. Its origin is usually said to be the work of Rice in 1894–1897. It was, however, a decade later before any continuous and general activity along this line began. Since then intelligence tests, at first individual and later group, achievement tests, and many other varieties of educational measuring instruments have developed rapidly and become very numerous. Chief among the causes of this rapid development were the published evidence as to the subjectivity of school marks, the school survey movement, and the publicity afforded by educational periodicals and other agencies. The first book definitely in this field is dated 1904, but it was not until about 1915 or 1916 that others in any number began to appear. Since that time scores of volumes having some connection with this movement have been published. Among the developments within the last ten or twelve years are general survey tests, the new examination, the spread of the movement from the elementary school into the high school and college, the development of a critical attitude toward tests, the measurement of character and personality, and other more or less minor ones too numerous to mention. At the present time there are available about fifteen hundred different standardized

or near-standardized tests or series of tests, and probably at least forty million copies are being employed annually.

References

Ayres, L. P. "History and Present Status of Educational Measurements," *Seventeenth Yearbook of the National Society for the Study of Education*, Part II. Bloomington, Illinois: Public School Publishing Company, 1918, Chapter I.

Freeman, F. N. *Mental Tests: Their History, Principles and Applications.* Boston: Houghton Mifflin Co., 1926, Chapters I–VI.

Kelley, T. L. "Historical Survey of Mental Measurement," *Interpretation of Educational Measurements.* Yonkers: World Book Company, 1927, Chapter I.

Levine, A. J. and Marks, Louis. "The Testing Movement," *Testing Intelligence and Achievement.* New York: The Macmillan Company, 1928, Chapter I.

Monroe, W. S. "The Beginnings of Standardized Objective Tests," *An Introduction to the Theory of Educational Measurements.* Boston: Houghton Mifflin Company, 1923, Chapter I.

Peterson, Joseph. *Early Conceptions and Tests of Intelligence.* Yonkers: World Book Company, 1925. 320 p.

Smith, H. L. and Wright, W. W. "Introduction," *Tests and Measurements.* New York: Silver, Burdett and Company, 1928, Chapter I.

Trabue, M. R. "How Standard Tests Developed," *Measuring Results in Education.* New York: American Book Company, 1924, Chapter III.

CHAPTER III

CRITERIA FOR THE SELECTION OF TESTS

Criteria to be considered.—It is the purpose of this chapter to present a list of criteria which should be employed in selecting standardized tests and to discuss and explain the meaning and application of each. Several of the same criteria are also suitable for use in connection with teacher-made or unstandardized tests, but others of them have no application in connection with such tests. The reader will have no difficulty in recognizing which are applicable to other tests and which are not, since the discussion without saying so explicitly will make clear that certain of them cannot be expected to apply to ordinary classroom tests.

The chief criteria or bases of selection, some of which may be subdivided, are as follows:

Validity
Reliability
Objectivity
Norms and other provisions for the use of results
Duplicate and equivalent forms
Scaling
Ease of administration
Cost

These are arranged in approximate order of importance with regard to the real merit of tests. From the more practical standpoint of selecting tests which it is possible or practicable to employ, the last two items are often of greater importance than their position indicates.

Validity.—The prime consideration in selecting a test should be its validity. A measuring instrument is said to possess validity if it accomplishes the purpose it is stated or intended to perform. For example, a foot ruler is a valid instrument for measuring height or length, but not for weight. Therefore,

if it were employed to measure weight, it would not have validity. To give another example, a written test in which directions and responses are in the English language is not valid for measuring any ability or trait of persons who do not understand English, although it may be highly valid for the purpose intended in the case of those familiar with the language. It is impossible to determine whether or not a test is valid unless one knows or assumes the function for which it is intended. In other words, a test is not valid in general, but only valid with respect to one or more specific functions.

Because of the fact just stated, it is highly desirable that the author or publisher of a test state its purpose so that one who is thinking of employing it may judge whether or not it is valid for that purpose. Occasionally a test is found to have such high validity for some other purpose than that for which it was intended as to warrant its use therefor. Furthermore, the statement of function should be specific rather than general. Thus, instead of merely announcing a test as being intended to measure knowledge of Latin, it would be better to state, for example, that it is intended to measure knowledge of Latin vocabulary and ability to translate Latin into English. It would be still better to go into more detail and state that it was intended to measure knowledge of the two thousand words most commonly employed in high-school Latin, and ability to translate into English sentences composed of those words and the grammatical constructions most commonly employed in high-school Latin. If, on the other hand, the purpose of the test is general, that is, if it is intended to yield a general measure of pupils' achievements in all of the important phases of Latin, the title or the announcement should so state. Unfortunately in many cases no such explicit statements of the supposed functions of tests accompany them. In only a minority of cases, indeed, are they contained in the titles of tests, but frequently by looking in the teachers' manuals or elsewhere, more complete statements thereof can be found.

Test validity is of two kinds, curricular and statistical. Since these are rather distinct from one another, each will be discussed in a separate section.

Curricular validity.—One method of determining the validity

of a test is to study its content from the standpoint of agreement with the particular curriculum in connection with which it is to be used, or with a curriculum that is generally considered to be of high merit. In so far as the determination of content is concerned, the same general principles that underlie curriculum construction apply to test construction. If it is to measure knowledge of what pupils achieve with regard to their opportunities to achieve, a subject-matter test should be based upon the curriculum employed and actual teaching procedure. On the other hand, if intended to measure achievement with regard to the best known, and perhaps also to exert some influence in determining what is taught, it should be founded on the best existing curriculum and teaching practice. There is a place for tests of both sorts. In the effort to insure curriculum validity of one sort or the other a number of different methods of choosing the content of tests have been employed. The chief of these will be treated briefly.

One method is to choose the material for the test on the basis of what is common to all or a number of the best or most typical courses of study. Frequently textbooks are used instead of courses of study. Sometimes the procedure is varied by employing topics most commonly dealt with in such examinations as those of the New York Regents or the College Entrance Board, or in a large number of examinations prepared by classroom teachers. Still another variation is to secure the opinions of teachers or supposed experts as to what should be included. Frequently two or more of these methods are combined. In general the results which they yield are similar and on comparatively few items is there very marked disagreement. All tend to produce tests that measure what is most generally taught or most commonly considered of importance. When the content of a test is based upon what a comparatively few experts believe it should be, the result is likely to be that it includes not what is most commonly taught but rather what should be taught. In other words, it tests the content of the course as well as pupils' mastery of that content.

Another procedure frequently followed in test making as well as in curriculum construction is the study of the supposed utility of the material included. Instead of determining how

often items occur in courses of study, examinations, experts' recommendations, and so forth, their frequency in so-called life situations is determined. Studies have been made, for example, of the words used or misspelled in written correspondence, of the historical dates and characters and the historical and geographical allusions found in literature of various kinds, of the mathematical principles and operations actually employed outside the school, of the grammatical constructions used and errors made, of the most common vocabulary in our own language and in others, and so forth. In most cases the construction of tests on this basis has been part of a larger project involving curriculum construction also. In so far as this is true, this method is not radically different from that previously described. In a few cases, however, measuring instruments have been constructed according to this procedure independently of any work on the curriculum. It should be observed, however, that when this has been the case and the measuring instruments have become at all popular, it has usually not been long until curriculum construction based upon the same or similar studies was under way, thus illustrating the principle that a widely used test influences the content of what is taught.

Before leaving this method of constructing either a test or a curriculum, the writer wishes to mention one fallacy sometimes overlooked. It has sometimes been assumed that such a compilation of practice outside the school yields a complete determination of what should be taught in school. If this were accepted and universally acted upon, progress would be limited, since each generation would study only what the previous generation was actually using. Indeed, not only would progress be limited, but probably even decline would ensue, since no generation makes use of all that it studied in school. In most, and perhaps in all, subjects, such methods yield only the minimum essentials of a course, and not always those. The words, grammatical constructions, mathematical operations, historical dates and characters, and so forth, common in daily life should, of course, be taught in school, unless they are thoroughly learned outside, but what is taught should not be limited to them.

With regard to curricular validity, a distinction may be made between general tests and diagnostic tests. A general test over any

body of subject-matter should deal with all, or at least a number, of the most important phases or parts thereof and also contain a well chosen sampling of the details. A diagnostic test, on the other hand, should ordinarily not attempt to cover the whole subject or any large portion thereof, but should contain a large number of items on some one or few limited divisions of the subject.

In the case of practically all worth while standardized tests, information concerning their derivation and construction which will be helpful in determining their curricular validity is made available either at or soon after the time of publication. Frequently it is contained in the teachers' manuals, also often in articles in educational periodicals. It is not always as detailed as would be desirable, but nevertheless worth consulting. From it and direct study of the test itself the teacher who is familiar with the chief curricular studies in a given subject and the general trend of development therein should be able to make a fairly intelligent selection on this basis.

Statistical validity.—Although there is some overlapping between this and curricular validity, yet a very definite distinction between the two can be made. Statistical validity has to do primarily with the selection and arrangement of content for a test so that the results obtained from giving it to pupils indicate it is accomplishing its desired purpose. A test may be highly valid in so far as the choice of items is concerned, but because of the form in which they are put, be very unsatisfactory, or vice versa. For example, a general test may deal with the hundred facts or ideas found most frequently in American history texts, but may present them in exercises almost entirely too difficult for high-school pupils to understand, so easy that practically all high-school pupils can answer them correctly, or in some other way unsatisfactory. On the other hand, a test may consist of items which behave exactly as desired statistically, but which because of their triviality or careless selection do not constitute a satisfactory sampling of what it is desired to measure.

The kind of items most desirable from the statistical standpoint depends to a considerable extent upon the type of test, especially whether it is a rate or a scaled test. The test elements for the former should not vary a great deal in difficulty, whereas for a scaled test they should range all the way from very easy to very

hard. Neither sort of a test should contain many items on which all or even nearly all pupils fail, nor many that all or nearly all get correct. Furthermore, not only should the items and the types of exercise in which they are included be such that pupils tend to make higher scores the longer they have studied the subject dealt with, but also such that each individual item obeys this same law. For example, in a foreign language test designed to be used through several years, it is ordinarily not desirable to include words used in first-year books but not later upon which pupils at the end of the first year would probably score better than those at the end of the second or succeeding years.

Though it is not the purpose of the writer to discuss test construction, nevertheless it seems in place to point out briefly what should be done to secure such statistical validity as has just been described. If he knows this, a teacher can often reach more satisfactory conclusions as to the validity of tests by comparing the steps in their construction with those which should be followed. A test maker should prepare two or three times as many items as he expects to use in his completed work, all of which have curricular validity. The exercises containing these items should then be tried out upon a sufficiently large number of pupils to yield reliable results and such eliminations, revisions, and rearrangements made as the results obtained indicate are needed. A second tryout should follow, and more if needed. Sometimes directions and forms of exercises need changing as well as items themselves.

It was stated above that there is some overlapping between curricular and statistical validity. The application of statistical methods, especially correlation,[1] to determine whether or not a test fulfills its given function, is commonly included under the head of statistical validity, but its purpose is really to determine curricular validity. The general procedure in this respect is to compare or correlate the scores yielded by the test in question with an independent criterion or, in other words, with another measure of the same thing which has no connection with the test itself. In the case of subject-matter tests, the two criteria most frequently used are teachers' marks in the school subjects and scores on other tests therein. Neither teachers' marks nor other

[1] Correlation is the agreement, or lack of it, between two series of paired scores. For a more complete explanation, see p. 580.

test scores are themselves perfectly valid, but since there is ordinarily nothing better available for the purpose, they are commonly employed.

On the whole the evidence secured from such comparisons or correlations has more negative than positive value. Low correlations between test scores and teachers' marks or other test scores in the same subject at least raise the question of whether or not the tests measure what they are supposed to measure, whereas high correlations with the same criteria are not proof that the tests in question really do fulfill their functions, although they tend to indicate this fact. If two or more general tests on first-year algebra, for example, constructed by different persons working independently, yield results which agree rather closely, it seems more likely that they are all measuring the same thing, and if this is announced as being general knowledge of first year high-school algebra, that this is what they are really measuring, than if the correlations between them were low. If the correlations are low, however, it is fairly certain evidence that they are not measuring nearly the same thing and, therefore, that one or more of them is not valid. In the case of group intelligence tests, a very common criterion is the Stanford Revision of the Binet-Simon Tests.[2] This is used because it is generally considered the best individual intelligence scale in the English language and an individual scale is, other things being equal, more valid than a group one.[3] In some cases attempts have been made to determine the validity of tests by comparing the scores thereon with various other criteria than those mentioned above, but usually the validity of these criteria has been so doubtful or the method of determining them so unreliable that the results obtained have little value.

Reliability.—Reliability is synonymous with accuracy. The reliability of a test may be defined as the degree to which a second application yields scores equivalent to those obtained from the first application. A second application may refer either to giving the identical test again or to using a duplicate and equivalent

[2] See p. 395.

[3] The reason for this is that by using an individual scale an examiner can test a pupil more thoroughly and intensively, noting all the details of his responses, than is possible with a group scale.

form [4] thereof. It should be noted that this definition does not state that the same scores must be yielded by the second application as by the first, but only that they must be equivalent or, in other words, that a uniform known relationship exist between those obtained at the two testing periods. For example, if a group of pupils takes one algebra test containing ten rather difficult examples and another similar one containing twenty relatively easy ones, and each pupil solves correctly just twice as many examples on the second test as on the first, the results would be said to have perfect reliability. If one knew the score made by a pupil on either test, he could determine exactly what the same pupil made on the other. The same would also be true if each pupil solved, for example, eight more examples on the second than on the first.

If a test is not reliable, it cannot be valid, since if a test does not measure whatever it measures accurately, it cannot measure the thing it is supposed to measure exactly. It is, however, possible for a test to be highly reliable and yet possess little validity. For example, a test supposed to measure general historical knowledge might be highly reliable, but because it required the reading of rather long or difficult exercises or questions, measure reading ability rather than knowledge of history. To give a second example, a test in the solution of simple algebraic equations such as $2x = 6$, $3y = 12$, and so on, might be largely a test of the speed at which pupils could write figures rather than of their ability to solve equations, yet at the same time be highly reliable.

The reliability of a test cannot be determined by examining the test itself, although inferences in regard to it can be made from such an inspection. The test must actually be tried out to yield the desired information. Fortunately it is becoming an established custom for the authors of standardized tests to secure data on reliability before tests are placed on the market and for the publishers to include these data in the teachers' manual or other supplementary material accompanying the test. In many instances such data may be found in periodical articles, monographs, or other publications, sometimes written by the authors of the tests

[4] For an explanation of what is meant by a duplicate and equivalent form, see p. 74.

concerned, sometimes by others. Also some of the books more or less similar to this volume which treat various standardized tests present evidence on this point. If figures as to reliability are not given in the manual or elsewhere, it is likely either that the test is new and they have not yet been computed, or that it has not been made by an expert. In general it is best not to employ tests for which no reliability data have been published.

The ordinary method of determining the reliability of a test is to compare or correlate results from two duplicate and equivalent forms given to the same individuals or perhaps results from employing the same form twice. Sometimes, especially if there is only one form of the test, separate scores computed upon the odd and the even-numbered items of the test, are compared. That is to say, items 1, 3, 5, 7, and so on are considered as constituting one form of the test and items 2, 4, 6, 8, and so on as another. The results yielded directly by such comparisons must, of course, be modified [5] to be equivalent to what they would be if the more usual method of comparing results from two forms had been employed.

From the standpoint of the person choosing tests for actual use, the most important point connected with reliability is probably that of how it is measured and what is the significance of reliability measures of various sizes. In the first place, whether or not the reliability of a test is to be considered satisfactory depends largely on the purpose for which the results are to be employed. Many tests that do not yield individual scores reliable enough to be trusted, do give fairly accurate average scores for groups of pupils of ordinary class size or larger. In other words, they are reliable enough for use in judging the work of a class or of its teacher, but not for that of individual pupils. Furthermore, high reliability is more requisite for diagnostic use than for general survey purposes. A test which includes only a few items from each of a number of phases or divisions of a subject may yield a single total score that is reasonably satisfactory for measuring a pupil's general achievement and determining his promotion or classification, but not for diagnosing his difficulties in particular topics or phases of the work.

There are four measures of reliability that are commonly em-

[5] See p. 592.

ployed in connection with standardized tests. These are the coefficient of reliability (r), the standard or probable error of measurement ($\sigma_{meas.}$ or P. E. $_{meas.}$), the ratio of this error to the mean (M), and the ratio to the standard deviation (σ). The methods of computing these will be found in Chapter XXIV, but a brief definition and explanation of each will be given here.

The coefficient of reliability.—The coefficient of reliability, also called the coefficient of self-correlation, is the coefficient of correlation between scores yielded by two forms of the same test. It ranges from + 1.00, which denotes perfect positive agreement, through zero, which denotes chance or no agreement, to — 1.00, which denotes perfect negative agreement between two series of scores or other measures of the same individuals. For example, if the weights of a number of pupils of various sizes were found at a certain time and again a month later, and the two series were correlated, the result would be + 1.00 if the increases had all been the same either in absolute amount as, for example, two pounds, or in proportionate amount as, for example, one-fiftieth of the weight at the initial period. If, however, the increases had not all been the same but had not differed markedly, the coefficient of correlation would still be positive and rather high, that is, not far below 1.00. If, however, the ones who were heaviest at the first weighing had lost weight and those who were lightest had gained weight, the correlation would be still lower, being negative if the losses and gains were so great that those who at first were heaviest were at the second period lightest.

Although the coefficient of reliability is probably the most frequently given measure of reliability, it is not very satisfactory because its interpretation depends largely on the range of ability in the group tested. For example, if the same test is given to pupils in all four years of high-school Latin, the resulting coefficient of reliability is ordinarily considerably larger than if it is given only to those in first-year Latin although, of course, the real reliability of the test is not different in the two situations. If certain other data are available, it is possible to correct or change coefficients of reliability so that they will all be upon the same basis. Since these other data are not always available, however, and since the formula which must be used presents considerable difficulty to those who are not mathematically trained,

this correction is not commonly made. Instead the other measures named are employed.

Errors of measurement.—An error of measurement may be defined as the difference between the score actually obtained by an individual upon a test and the score that accurately indicates the amount he possesses of the ability in question. This latter is commonly known as the theoretically true score, the word "theoretically" being inserted because it cannot actually be determined. It is impossible to determine the error of measurement in any particular score, but the errors in any group of scores can be described fairly accurately. They are most frequently measured and reported in terms of the standard or the probable error of measurement.

The standard error of measurement is of such a size that it includes slightly more than two-thirds [6] of the errors of measurement. For example, if the standard error of measurement of a particular test is four points it means that slightly more than two-thirds of the errors or, in other words, the errors present in the scores of slightly more than two-thirds of the individuals who took the test, are not greater than four points, whereas the remainder, slightly less than one-third, are four points or larger. Since, however, the proportion of errors included by the standard error is not a round number, it has become more common to employ instead the probable error of measurement,[7] which is similar to the standard error except that it includes just half of the errors. For example, if the probable error of measurement of a group of scores is three points, half of the scores are not in error by more than that amount, whereas the other half are in error by three points or more.

The interpretation of either the probable or standard error may go further than has just been suggested. A statement can be made as to the limit or size of error which includes any desired proportion of the errors. Ordinarily this is done in terms of two, three, or some other even number of standard or probable errors. Thus not only are half of the errors in a group of scores not greater

[6] The proportion included by the standard error is more nearly 68.27 per cent. It is merely the standard deviation (see p. 63) of the errors.

[7] The probable error of measurement is simply the median deviation (see p. 578) of the errors.

than the probable error, but also about 82 per cent of them are not greater than twice the probable error, almost 96 per cent not greater than three times the probable error, more than 99 per cent not greater than four times the probable error, and so forth. Therefore if we know that the probable error of a particular set of scores is three points, as was assumed above, we not only know that half of the errors are not greater than three and the other half three or more, but also that about 82 per cent are not more than six points and 18 per cent six or more, that almost 96 per cent are no greater than nine and 4 per cent that large or larger, and so forth.

Ratios of errors of measurement to the mean and standard deviation.—It can readily be seen, however, that although a statement of the probable error of measurement yields helpful information as to the likelihood of an error of any given size in a particular score, yet it does not furnish a satisfactory basis for comparing the reliability of one test with that of another. Thus a probable error of three points seems large for a test on which the total possible score is only twenty points, whereas for one on which the total is one hundred and fifty points it does not appear to be of such great significance. Because of this it has been suggested that the probable error of measurement be compared with some other measure that is indicative of the size or range of scores on the test. For this purpose two have received rather common use. These are the mean [8] or, in other words, the ordinary arithmetic average, and the standard deviation.[9] Each has certain advantages over the other, so that it seems best to give both in a complete statement of reliability. The fact that pupils make the same mean score on two tests does not indicate necessarily that the distribution of scores on the two is the same. For example, on two tests of one hundred points each the mean scores might be sixty-five. In one case, however, all the scores might happen to be between sixty and seventy, whereas in the other they might range all the way from twenty to one hundred. Thus an error of measurement

[8] See p. 568.

[9] The standard deviation is a measure of the variability or spread of a group of scores that includes the 68.27 per cent of them which are nearest the average or, in other words, have the smallest variations or differences from the average. See p. 574 for the method of computing it.

of any given amount would be much more significant in the first case than in the second, since the total range of variability among the scores was much smaller. It has, therefore, been urged that the divisor used be the standard deviation, which is a measure of the spread of variability of the whole group of scores. Thus the ratio of the probable error of measurement to the standard deviation of the whole group of scores indicates the size of the error of measurement with regard to the total range or variability of the scores or, in other words, in comparison with the possibility for variation or error. On the other hand, the comparison of the error of measurement with the mean appears to yield rather helpful information by indicating what proportion of the score the error is likely to be.

Interpretation of measures of reliability.—Although it is difficult to interpret reliability measures of various sizes by means of words or statements, Table I presents an attempt to do so. In this coefficients of reliability, ratios of the probable error of measurement to the mean and to the standard deviation[10] are divided into five groups according to size and an attempt made to state in words the significance of those in each group. Such interpretations are not very useful for the coefficient of reliability, because its size depends largely upon the range of ability of the group tested, but for the other two measures they are more helpful. Furthermore, as has been stated previously, reliability depends in part upon the length of time necessary to administer a test, therefore in interpreting the significance of a measure of reliability one should take into account the time limit of the test. The figures given in Table I may be taken as applying to tests for which the actual working time is about thirty-five or forty minutes. They are based on present standardized tests rather than on what would be desirable or ideal, so that the standard of judging is comparative rather than absolute.

The first line of Table I indicates that if the coefficient of reliability of a test is .95 or higher, the ratio of the probable error to the mean not over .03 and its ratio to the standard deviation not

[10] Since errors of measurement are stated in terms of the scoring system employed upon particular tests, and since such scoring systems vary greatly from test to test, no general statement can be made as to the interpretation of such errors.

TABLE I. SIGNIFICANCE OF COMMON MEASURES OF RELIABILITY OF VARIOUS SIZES WITH REGARD TO THE USE OF THE TESTS TO WHICH THEY APPLY

Coefficient of reliability	$\frac{\text{P. E. meas.}}{M}$	$\frac{\text{P. E. meas.}}{\sigma}$	*Per cent of tests included* *	*Significance*
.95–1.00	.00–.03	.00–.15	10	Among the very best. Only a few tests have reliability this high.
.90– .94	.04–.05	.16–.22	20	Comparatively high. The number of tests in this group is increasing rapidly, but is still much smaller than the number below it.
.80– .89	.06–.08	.23–.30	35	Satisfactory for group measurement,† but only fairly so for individuals.
.70– .79	.09–.11	.31–.37	20	Fairly satisfactory for group measurement, but should rarely if ever be employed for individuals.
.00– .69	.12 or more	.38 or more	15	Should never be used alone for individual ratings and rarely for small groups.

* These figures, which are only approximate, are based on the tests described in this book for which data are available.

† The reliability of an average score for a group increases in proportion to the square root of the number of persons or scores concerned. That is, the average score of a group of sixteen pupils is four times as reliable as an individual score, for a group of twenty-five pupils it is five times as reliable, and so forth. In other words, a test on which the ratio of the probable error of measurement to the mean equals .20 for an individual score has a similar ratio of only .04 for the average score of a class of twenty-five pupils, and of only .05 for a class of sixteen.

over .15, the test is among the less than 10 per cent that possess such high reliability and therefore deserves very high rank in this regard.[11] The second line shows that tests with coefficients of

[11] This high rank is justified rather because of the fact that few tests possess such high reliability than because reliability measures of the size indicated are in themselves nearly perfect. A fuller interpretation and explanation of the meaning of coefficients of correlation or of reliability of various sizes is given on page 595.

from .90 to .94 and ratios of .04–.05 and .16–.22 respectively, rank in the upper 30 per cent of all. The other lines of the table may be read in a similar way. There are coming to be more and more tests that yield measures within the limits of the two highest groups. It has sometimes been suggested that .90 be accepted as a standard for the coefficient of reliability which all tests should attain. No such exact critical point can be justified, but it is probable that within a few years a majority of the tests receiving wide use will have reliability at least this high.

The reader may wonder why any tests falling in the lowest group are described in this volume or in any way recommended. There are two explanations for this. One is that they are in fields wherein the number published is so limited that there is little choice. The second is that although they are decidedly low in reliability, they rank high enough in the light of other criteria to merit use, at least for the measurement of groups. In this connection the reader should keep in mind that the reliability figures actually given in connection with descriptions in this book and elsewhere are for the tests with their actual time limits, which in some cases are decidedly shorter than thirty-five or forty minutes, and would be raised considerably if the tests were lengthened. This point will be discussed further in the following paragraph.

Factors that affect reliability.—The reliability of a test depends on various factors. One of the most important of these is its length. Both from the standpoint of the content of the test itself and of the reaction of the pupils, it is in accord with common experience and proven by actual experimentation that if the length of a test is increased up to a reasonable limit,[12] its reliability is increased. This results because the greater the number of items included the better the sampling of what it is desired to measure and therefore the pupils' responses are more representative of their true abilities. It has been shown for similar tests, that is, tests covering the same subject or phase of a subject and containing the same types of exercises, the reliability of a pupil's

[12] A reasonable limit may be defined as one not so long that fatigue and other disturbing factors seriously affect pupils' work.

scores increases approximately [18] as the square root of the increase in length. In other words, if one of two similar tests is twice as long as the other, the score is approximately 1.4 times as reliable; if it is four times at long, it is about twice as reliable, and so on.

Another factor that influences reliability is the scaling or arrangement with respect to difficulty of the items or exercises in the test. A test that does not have a large number of very easy or very difficult items but has more items near the middle range of the ability of the group to be tested is, other things being equal, more reliable than one which does not contain such items. Also a test which is scaled in finer units tends to be more reliable. This is analogous to the situation that would exist in measuring height if one measuring instrument contained no divisions finer than inches whereas another was graduated to sixteenths of an inch. It can readily be seen that more reliable or accurate measurements could be obtained with the second or finer instrument.

The reliability of a test also depends to a certain extent upon the wording of the directions to pupils and of the test exercises themselves. In most of the standardized tests which have appeared within the past few years the wording is sufficiently satisfactory that there is very little loss of reliability due to this cause. There are, however, enough exceptions to this that anyone selecting a test should examine critically the directions as well as the test itself from this standpoint. Among the points which the directions should provide for are the following:

State briefly what the test is about.

Instruct pupils when to begin and when to stop work, when to turn a page or not to turn a page, and so forth.

Direct pupils whether to delay on each item until they have answered it or to go ahead if they do not know it.

Make clear the form of recording answers, whether by writing words or numbers, underlining, checking, or something else, in-

[18] The increase in reliability is only approximately equal to the square root of the ratio of the lengths because the two tests concerned are never exactly similar.

cluding a fore-exercise [14] to illustrate the method of response unless pupils are already thoroughly familiar with it.

Tell pupils what to do in case they reach the end before time is called.

The conditions under which tests are given also exercise some influence on the equivalence of scores although if reasonable precautions are taken this appears to be slight. For example, results will probably agree slightly more closely if the two forms of a test are given at the same time of day and perhaps even if given on the same day of the week, but in most situations this is not very important. It will, however, be liable to lessen agreement seriously if one form of a test is given at the middle of the morning of an ordinary school day and another form the last thing in the afternoon immediately before some social or athletic event in which pupils are very much interested and for which they are very eager to get out of school. Also it tends to cause discrepancies if the two forms of a test are given by different individuals, although if directions are adhered to as closely as they should be this can usually be neglected. The shorter the period of time between the giving of two forms ordinarily the greater the agreement, but here again the difference is commonly small unless something helpful in connection with the test has been learned meanwhile. It is especially true in the case of intelligence tests that the lapse of considerable periods of time has little influence in lowering reliability. Another condition that should be considered is the familiarity of the pupils with standardized tests, or at least with tests of the type used. If they do not possess such familiarity scores at the first testing will in all probability not be as representative of their ability as those secured later. Therefore the agreement between first and second testing scores will ordinarily not be so high as that between the latter and those from a third testing. Indeed, some writers have advocated that scores on the second instead of the first administered form of a test be taken regularly as measures of pupils' ability since they are more reliable than those obtained on the first form, and that the first form

[14] A fore-exercise or practice exercise is one similar in form to the test itself to which pupils are to respond to make sure that they understand what is to be done. It does not count toward the score.

be given rather to prepare pupils for the second than to yield usable scores.

Constant and variable errors.—In connection with reliability mention should be made of constant and variable errors. A constant or systematic error, in the strict sense of the term, is one that is the same for all of a given group of pupils. Since errors that fulfill this condition exactly rarely occur in test scores, the term is commonly interpreted more loosely to include any error that tends to be the same, either absolutely or relatively, for all members of a given group. For example, if a teacher accidentally allows too much or too little time for a test, the resulting errors are considered constant since in the first case all pupils have the same amount of time wherein to increase their scores and in the second lose the same amount. An error that is absolutely constant for a group of pupils at one testing period but not present at another does not affect the correlation between the two series of scores, hence it does not lower the reliability of the results. In general, therefore, constant errors may be neglected in so far as their effect upon reliability is concerned.

Variable errors, on the other hand, are those which differ for the various individuals composing a group. For example, if a pupil breaks a pencil point and loses a little time, if he happens to see his neighbor's paper and learns what the correct answer to an exercise is, if a pupil just happens to have studied certain of the items contained in a test or to have omitted them in his study, if he happens to be in the very best physical and mental condition or to be decidedly below par, variable errors result. Such errors lower the reliability of the scores made. If, however, proper precautions are taken, the effect of most of these errors can be reduced to a minimum. For example, pupils should have ready two pencils so that no time will be lost if one is broken; also pupils who are evidently in very poor physical or mental condition should not be allowed to take tests at that time. In general if pupils are in good enough health to justify their being present at school, this factor causes only slight errors in test results. This is especially true if, as is generally the case, tests are short enough that the point of serious fatigue is not reached.

Objectivity.—Just as reliability is a necessary factor in valid-

ity, though the reverse is not true, so objectivity is vital to reliability although not so inclusive. The term objective is commonly applied to a test which contains exercises of such a sort that there is no disagreement among competent scorers as to what the correct answers are.[15] In other words, objectivity is just the opposite of subjectivity in scoring and is the quality which results from the elimination of scorers' judgments. It is perhaps needless to point out how this contributes to reliability. Certainly no test can be expected to give reliable results from time to time if there is disagreement or doubt as to what the correct answers are, so that at one time certain responses are considered correct whereas at another time others may be so considered.

In contrast to reliability, objectivity can be largely determined by an inspection of the test itself, including the scoring key. The key should, if possible, include all answers which are to receive credit as correct and definitely state that no other answers are to be accepted. In most cases it is possible to make tests so that there is one and only one correct answer to each element. This not only insures perfect objectivity but likewise renders scoring easier. Moreover, the exercises should be so formulated that there can be no argument as to the correctness of the key. In many cases the published scoring keys do not attain this desired perfection. Instead they give a list of correct answers and allow others which appear to be correct to be accepted. Sometimes they give lists of the wrong answers found to occur most commonly so as to warn teachers against accepting them. In a few cases also they contain doubtful answers which are to receive partial credit.

Not only should the scoring key contain a complete list of correct answers but there should be exact and detailed directions for computing the scores. These directions should cover what to do in the case of such points as erasures or alterations by pupils, illegible handwriting, misspelling, the use or omission of such symbols as the dollar sign, the per cent sign, and so forth in

[15] The writer realizes that this definition is more limited than that frequently employed, especially in the physical sciences. In a more complete sense objectivity includes not only the concept stated above, but also that all competent scorers are agreed as to the relative value or weight to be given each response. This is a condition which never, or practically never, holds true in regard to educational tests.

mathematics, whether or not mathematical answers must be reduced to lowest terms to be considered correct, and so on. Unless all such points are provided for, scorers will disagree as to whether or not to allow credit.

It should be pointed out in this connection that there are certain types of educational measuring instruments which because of their very nature cannot be made perfectly objective. The most common of these are such quality scales as are used for measuring composition, drawing, handwriting, and so forth, and rating scales for pupils' traits and characteristics. Such measurements as these involve the use of scales rather than of tests. These scales are composed of series of specimens or descriptions with which pupils' performances, abilities, and traits are to be compared. Thus absolute right or wrong is not involved. For example, it cannot be said that one pupil's handwriting is exactly correct and another's entirely incorrect. Therefore in all such cases judgment must be exercised by the scorer. The objectivity of such instruments depends largely upon the suggested conditions under which such judgments are to be exercised, the provisions and directions to assist those making them to do so on the same basis, whether or not the judgments called for are such as can probably be made with reasonable accuracy by the persons who are to make them, and so on. Furthermore, it has been shown that the objectivity and, therefore, the reliability, of such judgments is ordinarily increased by intelligent practice in the use of such instruments.

Norms and other provisions for the use of results.—A very important question in connection with standardized tests is that of what sort of norms are available. This includes the kind of group or groups from which the norms were secured, the number of groups, and how the norms were derived. For some tests the only norms available are general or nation-wide ones. Although general norms are valuable and helpful, they are not sufficient. It is desirable that other norms of more limited scope be available to afford more appropriate bases of comparison for particular pupils or groups of pupils. It is not possible to give a complete list of the varieties of such norms, but as illustrative of possibilities along this line and also of what is actually being done in the case of some tests, the following may be named: state norms;

city norms; norms for city and rural schools; norms for systems of different sizes; norms according to the plan of organization, whether 8–4, 6–3–3, 7–5, or some other; norms for pupils of different races and nationalities, of different degrees of intelligence and so forth; norms for different textbooks or other variations in the same course; norms for first, second, third, and even later testings.

There are a number of different bases according to which norms are given. They are generally computed for each grade or half grade in elementary-school subjects, according to length of time studied in high-school subjects, and for each age in the case of intelligence. In addition to age, grade, and time norms, such other varieties as T-scores, C-scores, percentile norms, and so forth are sometimes employed in connection with high-school tests.[16]

The question naturally arises as to how many of these numerous types of norms should be provided in order to justify the high rating of a test on this basis. With regard to norms for different kinds of pupils, types of schools, localities, and so forth, the general answer is that although the more the better, comparatively few are provided for most tests, even the better ones. Ordinarily the test manual gives only general norms for different groups of pupils according to some one or more of the bases already mentioned. If a test has been used for some time, however, one may find statements of the results obtained from using it in one or more state, city or other limited testing programs. Unfortunately, however, most of such reports appear in bulletins and other publications of state, university, or city research bureaus or other departments not easily found by individuals who do not have access to a rather complete educational library and perhaps also considerable time available for searching. Therefore the classroom teacher will usually have to judge the worth of a test on this basis according to what is given in the manual.

With regard to the basis of determining norms, that which seems best for high school is the length of time a subject has been studied. This is not entirely satisfactory, however, since, other things being equal, the pupils in each high-school class are a

[16] Chapter XVIII contains a discussion of these and other types of scores.

slightly more select group and somewhat more mature than those in the class immediately below and, therefore, in the same length of time should do slightly more and better work. This difference is recognized in the norms provided for a few tests. Usually such norms are stated for the end of one semester, two semesters, three semesters, and so on, of study of the given subject regardless of whether it was begun in the freshman, sophomore, junior, or senior year. There are some subjects in which practice as to when they are studied is nearly enough unanimous that a single set of norms of this sort is satisfactory or approximately so, but for many others this is not the case and different sets of norms should be provided. For the present any test provided with norms for the end of each semester of study of a subject, according to the year in which begun, may well be considered satisfactory in this respect. For example, for pupils who begin Spanish as freshmen there should be norms for the end of each of eight semesters, for those who begin it as sophomores for the end of each of six, and so on.

Source of norms.—Another point to be considered in connection with norms is whether or not the selection of pupils to represent the groups to which the norms are intended to apply was satisfactory. In many cases those making tests have relied too largely on mere numbers and have not paid sufficient attention to the selection and distribution of such pupils. In some cases, for example, thousands or tens of thousands of pupils have been tested, but all were enrolled in the schools of one or a few large cities, or in a single state or perhaps a few states in the same part of the country, and general nation-wide norms announced on the basis of the results secured. In other cases the geographical distribution has been sufficiently wide and representative, but some other selective factor has influenced the results. Perhaps all of the schools were in large cities, or in small towns. In many, perhaps most, cases, norms are based upon the giving of tests to schools which volunteer for the purpose. On the whole it is probable that schools or teachers sufficiently interested in the educational measurement movement to volunteer for this purpose are somewhat above the average and, therefore, that norms based thereon are too high for pupils in general. Because of these and other possible sources of bias in the selection of the pupils tested

a teacher should if possible learn whence the norms were secured and the basis of selecting the schools or groups of pupils utilized.

Other provisions for the use of results.—Although the provision of satisfactory norms is the most important item in connection with using the results of most tests,[17] yet there are other points that should also be considered. This is especially true of tests to be employed for diagnostic purposes. If such tests include a number of parts covering different phases or portions of the subject, provision should be made for obtaining separate scores on the several parts. Furthermore such tests should be accompanied by material of various sorts to be used for diagnostic purposes. This material should include, for example, suggestions and, when obtainable, actual data as to the probable causes of errors of different kinds and also advice as to the best remedial methods for overcoming these errors. It is hardly to be expected that a teachers' manual or other material accompanying a test will cover the matter of remedial instruction completely, but several pages may well be devoted to this purpose.

Duplicate and equivalent forms.—The two expressions "duplicate form" and "equivalent form" are commonly used synonymously and in a sense which really embodies the meaning of both. In its exact meaning, a duplicate form is one exactly similar to the original form in types and numbers of exercises, whereas an equivalent form is one of exactly the same difficulty. Ordinarily, however, when either term is used it is implied that both of these conditions are met. Also the word "form" alone is often used with the same meaning. In the case of any test that one desires to repeat with the same pupils, it is advantageous if there are duplicate and equivalent forms. It is true that in a few instances it is possible to repeat identically the same test, especially if the time interval is fairly long, with satisfactory results, but both from the standpoint of avoiding practice effect in repetition and also from that of being able, if desired, to test more completely by giving both at one time, it is advisable to employ tests having two or more such forms.

Practically all of the best standardized tests as contrasted with

[17] The chief exceptions are practice tests and, to a lesser degree, diagnostic tests.

scales have at least two duplicate forms commonly designated A and B, I and II, or in some other similar fashion. Some tests have three forms, and a few four or even more. In almost all cases these different forms are exact duplicates as to the numbers and types of items and exercises, and in most cases their difficulty is nearly enough equivalent that for practical purposes no difference need be noted. In a very few cases in which the two or more forms are not almost equivalent, different norms are given for them or else the directions provide for adding or subtracting a certain amount or multiplying by a certain ratio to make the scores on the different forms equivalent.

In many cases the equivalence of the forms of a test is not so exact and thorough as it should be, but applies to average rather than to extreme scores. In other words, pupils who are close to average will make approximately equivalent scores on the different forms, but those who are decidedly above or below average may not. To illustrate this, suppose that Form A of a fifty-element [18] test contains twenty very easy elements, ten of medium difficulty, and twenty very hard ones, whereas Form B contains twenty only slightly easier than average, ten of medium difficulty and twenty only slightly harder. In general, pupils of average ability will be able to respond correctly to the first twenty elements on each test and to about the same number of the middle ten on each, but to few if any of the last twenty and thus earn practically the same average score on both tests. On the other hand, pupils of considerably below average ability may answer correctly most of the first twenty on the first test, but almost none of those on the second, whereas pupils of somewhat above average but not of very superior ability can do few if any of the twenty very difficult items of Form A, but practically all of the twenty somewhat difficult ones on Form B. Thus inferior pupils will make decidedly better scores on Form A than on

[18] The term "element" is frequently used to refer to each part of a test that calls for a separate response on the part of pupils. Thus a true-false statement or yes-no question, a multiple-answer question including its several suggested answers, one pair of items in a matching test, is each a test element. Sometimes the expression "item" is used in the same sense, but it is also sometimes employed somewhat differently. For example, it may refer to each of the suggested answers in the multiple-answer element, or to each word or other expression in each column of a matching test.

Form B, whereas for superior but not highly superior pupils the case will be reversed. Therefore for two forms to be satisfactorily equivalent not only should the average scores made by the same group of pupils thereon be the same, but also the spread or variability of the scores, that is, how far they are from the average. In other words, the distribution of scores made on the two forms should be the same at all points. For example, if 2 per cent of the pupils make scores above ninety on Form A, there should also be 2 per cent who make scores above ninety on Form B; if there are 5 per cent between eighty and ninety on Form A, there should also be 5 per cent within the same limits on Form B, and so on. Unfortunately specific information on this point is not commonly furnished in the supplementary material accompanying tests.

In this connection it seems well to describe very briefly the methods used to secure forms which are duplicate and equivalent, especially the latter, even though most readers will have no occasion to apply them. The best method involves planning for as many such forms as are desired from the beginning and, therefore, formulating and trying out a large enough number of items that after the unsatisfactory ones have been eliminated the others may be divided to compose the desired number of forms. In this division items should be paired if there are to be two forms; grouped in threes, if three; and so on, each pair or group being of the same difficulty and one of each being placed in each form. A slight variation of this which accomplishes practically the same thing is to arrange all the items to be used in order of difficulty and then if two forms are wanted place items 1, 4, 5, 8, 9, and so on in one form, and items 2, 3, 6, 7, 10, and so on in another; if three forms, items 1, 6, 7, 12, and so on in one, 2, 5, 8, 11, and so on in another and 3, 4, 9, 10, and so on in the third; and similarly for more than three forms.

In the case of tests of which one or more forms have already been constructed and others are desired, an ample supply of material of the same general sort as in the first should be prepared and tried out along with all or perhaps a part of that in the original test. Sufficient material of equal difficulty should then be chosen to make an equivalent form. Sometimes another form is prepared by taking the same items of knowledge or ability included in the original form and changing the wording or perhaps

the type of the exercise which deals with each. Thus items contained in multiple-answer exercises in the original may be placed in matching, true-false, or completion in the second form, those in completion form in the first in matching, true-false, or multiple-answer in the second, and so on. After this has been done it is, of course, necessary to try out the two forms to insure that they are equivalent and ordinarily to make some alterations or adjustments in the second form before this result is attained. It is generally considered undesirable to construct a duplicate test in this way, however, except perhaps in cases in which the supply of good items is so small that if additional ones are selected they are liable to be trivial or otherwise unsatisfactory. One reason is that if the same items are covered by the two or more forms it is not possible to use them together to secure a wider or more complete measure of pupils' ability or performance. Also the practice effect when the second form is given tends to be larger than if different content has been employed. There are few subjects in which this limitation of items is so severe that it is not easily possible to prepare two satisfactory forms of a forty-minute test over any semester or year's work, or even any major division thereof. In some cases, such as intelligence, vocabulary, algebra, and so forth, the supply is for all practical purposes unlimited.

Scaling.—Although some test experts emphasize the importance of having the items of a test scaled, that is, arranged in order of increasing difficulty, the writer does not believe that this is in all cases vital. Rather does its necessity or desirability depend upon the function to be served by the test. As was suggested in the discussion of validity, scaling and the determination of difficulty has frequently received too great attention to the neglect of the proper choice of content. On the other hand, there are some definite advantages to scaling. As has already been indicated, exact determination of the difficulty of test items is practically necessary in the construction of equivalent forms. Furthermore, whether a test is primarily a scaled test covering a wide range of difficulty or a rate test covering a comparatively small range, there is probably some advantage in arranging the exercises or items in the order of difficulty. For a scaled test this is implied and is imperative if the test is to function as it is commonly assumed such a

test should, that is, permit pupils to solve all the examples within their power and to be stopped only by those that are too difficult for them rather than by lack of time. On the other hand, in a rate test in which the time is so limited as to make speed a real factor it is at least sometimes desirable that all pupils, whether they come near to completing the test or not, have the opportunity of attempting some of the relatively easy and some of the relatively difficult elements thereof, which would not be the case if they were arranged in strict scalar order. For certain experimental purposes scaling is necessary to secure the data desired, data often not of great importance in regular classroom testing.

If a test is scaled there are certain general principles that should ordinarily be followed. Usually such a test should begin with one or a few elements so easy that every or practically every member of the group to be tested can respond correctly and should progress by as regular and equal steps as possible up to one or a few items so difficult that no or practically no members of the group can respond to them correctly. The difficulty of the items included should usually be such that the average score made is about one-half of the maximum. Furthermore the increases in difficulty from element to element should be relatively small and not great enough to cause a considerable accumulation of undistributed scores or a bunching of scores at any one point. If, for example, the first five items of a test have difficulties proportional to the numbers 1, 2, 3, 4, and 5, whereas the sixth item has a difficulty proportional to 10, there is likely to be a considerable bunching of scores at 5 including both pupils who were just able to answer element 5 correctly and also those who could answer elements of difficulty 6, 7, 8, and 9 if they were included, but not 10. Thus the score earned would make no discrimination between the abilities of pupils able to do items of the various degrees of difficulty from 5 to 9.

The weighting of scores.—One point which arises in connection with scaling is whether or not weighted scores shall be used. It was formerly much more common than now to employ the exact or approximate weights of the various elements of a test as scores, thus requiring the person computing the score to add a number of figures of different sizes. It has now come to be the general practice to count the same number of points, usually

one, on each element, thus reducing the labor in scoring very considerably. It has been found that for fairly large numbers of items the correlations of unequally weighted scores with those determined by counting the same number of points, generally one, on each item are so high that the gain in accuracy resulting from using weights is usually not worth the extra labor required to compute it. This is especially true if the items are arranged in scalar order with not too extreme or irregular variations in differences. In addition to this fact, however, there is another argument against using scores weighted on the basis of difficulty. The fact that one item in a test is more difficult than another is no justification for giving a pupil greater credit for answering it, thus implying that it is much more important than the other item. Consider, for example, two such questions as "What was the date of the Declaration of Independence?" and "What was the date of the treaty of peace with Mexico?" On the basis of difficulty for an ordinary group of pupils or other individuals the second question is many times as difficult as the first. Certainly, however, no one would maintain that the latter date is many times as important as the former for the pupil in American history to know and, therefore, that many times as much credit should be given for knowing it as for knowing the date of the Declaration of Independence. The writer recommends, therefore, for both theoretical and practical reasons, that when other things are equal tests be selected which do not make use of weighted values in determining the scores.

Ease of administration.—From what might be called the theoretical or scientific standpoint, that of the real merit of a test, the ease with which it can be administered is not of high importance. From the practical standpoint, however, it is decidedly important. Furthermore, it is possible in most school subjects and other fields in which a teacher may wish to employ tests to prepare tests that are fairly easy to administer without sacrificing other desirable qualities. The person selecting tests is, therefore, justified in paying attention to this feature.

Probably the prime factor that makes for ease of administration is that the directions for giving and scoring the test be fairly simple and easily understood, but adequate. They should make clear to both pupils and teachers just what is to be done and how

it is to be done, but should not be unnecessarily long. If the test is divided into several subtests or parts, each should have its own brief directions immediately preceding it in addition to such general directions for the whole test as are necessary. Fore-exercises should be included at the beginning of each test that is composed of a type of exercise with which pupils may not be thoroughly familiar. The test booklet which is placed in the hands of the pupils should not contain the directions intended for the teacher, but only those for the pupils. Including the former merely tends to delay or confuse pupils.

The provisions and directions for scoring should also be such as to simplify the procedure and reduce the labor thereof as much as possible. For most tests keys can and should be provided with answers so arranged that they can be laid alongside a pupil's responses. Ordinarily scoring is facilitated if the test exercises provide for pupils recording their responses in a straight column. Furthermore, the method of computing scores after the responses have been marked should be as simple and direct as possible. As mentioned in a previous section, the counting of each item or element as one point makes scoring much more rapid than if varying weights are given. If weights are given, however, they should be in whole numbers, not in fractions.

Another factor that should perhaps be mentioned under ease of administration is that of derived or transmuted scores,[19] that is, of scores changed from their original form to some other basis. In connection with many tests provision is made for derived scores of some sort or other. Indeed a number of tests make no provision for using or interpreting the original point scores, but require that they be transmuted into derived scores. This is never desirable. In some cases the actual calculation of derived scores is rather laborious, in others not. In all cases in which derived scores are to be used tables should be provided from which the derived score corresponding to any point score may be read off at once.

Most of our present-day standardized tests meet this criterion fairly well. Ordinarily the directions are reasonably short and simple and the provisions for scoring and method of computing scores such as to lighten the task as much as possible. There are,

[19] Derived or transmuted scores include age scores, quotient and ratio scores, *T*-scores, *C*-scores, grade or *B*-scores, and others. See Chapter XVIII.

however, some which do not conform approximately to the most desirable practices in this regard. For example, one test of considerably high merit in certain respects has scores weighted in such a way that fractions with eight different denominators enter into the scoring. It has been found that although only about five minutes per paper are required by trained clerks to mark a pupils' responses, about seven or eight are necessary to determine an individual's score after all his responses have been marked. If fractions were eliminated by multiplying all the weights by a certain number and perhaps changing some of them slightly, this time could easily be cut in half.

Cost.—Cost is, of course, a very practical and necessary consideration in connection with the use of standardized tests. It is, however, easily possible to overemphasize it. The total cost of even a relatively heavy testing program is such a small proportion of the total cost of instruction that one should not strive to limit expenditures for this purpose at the possible sacrifice of the quality of the tests used. Even if the more expensive tests are employed by all high-school teachers their total cost will probably not exceed 1 per cent of the total cost of high-school education. In most cases the cost of the whole testing program in a school is decidedly below this. On the other hand, there is no reason why one should pay any more for good tests than is necessary.

There are several reasons why the cost of tests that in many ways appear similar differs considerably. Of these probably the most important is that some publishers of tests spend much larger amounts than do others on expert editorial service, perhaps also to assist authors to construct and standardize their tests, and publish much more complete manuals and accompanying material. In other words, the buyer gets just about what he pays for. There is also some variation in the quality of paper used, the spacing, and so forth, but this is usually a relatively minor factor in determining cost. Some tests are put out by publishers who do very little business along this line and it is ordinarily necessary for them to charge somewhat more than large publishers of tests in order to make their business profitable. On the whole, one is more likely to be sure of getting his money's worth by purchasing a test from one of the few large publishers than by securing it

from some printing company or other concern which publishes only the one test or perhaps a very few.

In most subjects it is possible to secure first-class tests for not over five or six cents per pupil. In many cases the cost is lower, sometimes even only two cents or so. In the case of quality scales and similar instruments that are not to be placed in the hands of the pupils but of which only one copy is needed by a teacher, the cost ordinarily runs from five or ten cents a copy up to as much as a dollar or more. There are a very few tests which are decidedly expensive, costing fifty cents to a dollar or even more per pupil. Their number is so limited, however, and so many other measuring instruments are available that a well rounded testing program for practically any situation can be planned without including any of them.

One point that should frequently be noticed in connection with cost is how many copies of a test must be purchased. It has become rather common for publishers to put up tests in packages of certain sizes, of which twenty-five is the most common, and not to break such packages. If the number of pupils to be tested is considerably less than twenty-five, or whatever the number in a package is, if it is only slightly larger than the number in one package so that two packages are required, or some other similar condition with regard to the number prevails, many more copies of tests in such packages must be secured than are needed. Sometimes this can be avoided by two or more teachers or even two or more schools planning to use the same test or tests and purchasing their materials together. If this is impracticable, plans may be made far enough ahead to provide for using the surplus.

Test rating scales.—The writer has made no attempt to weight the various criteria or points coming under them in the preceding discussion. A few scales or lists of criteria which do so have been prepared by others, however. Among these is that of Otis, whose scale for rating tests embodies about the same criteria as have been discussed. It is given in Table II.

TABLE II. TEST RATING SCALE *

Manual (5)
Validity (15)

* From "Scale for Rating Tests," *Test Service Bulletin* No. 13. Yonkers: World Book Company. 6 p. By permission of the publishers.

Reliability (10)
Reputation (5)
Ease of Administration (Total 15)
  (a) Preparation (4)
  (b) Time limits (4)
  (c) Explanation needed (3)
  (d) Alternative forms (4)
Ease of Scoring (Total 15)
  (a) Objectivity (10)
  (b) Time required (3)
  (c) Simplicity (2)
Ease of Interpretation (Total 15)
  (a) Norms (5)
  (b) Directions for interpreting (4)
  (c) Class record (1)
  (d) Application of results (5)
Convenient Packages (5)
Typography and Makeup (5)
Test Service (10)
  Total (100)

Another somewhat similar scale, suggested by Cole and von Borgersrode, may be found in Table III. It contains many more

TABLE III. SCALE FOR RATING STANDARDIZED TESTS *

I. Preliminary Information
  1. Exact name of test.
  2. Name and position of author.
  3. Name of publisher and nearest address.
  4. Cost.
  5. Date of copyright.
  6. Purpose of test.
II. Validity (25)
  A. Curricular (15)
    1. Exact field or range of educational functions which test measures?
    2. Ages and grades for which intended?
    3. Criteria with which material was correlated?
    4. Do questions parallel good teaching procedures?
    5. How wide is sampling of important topics?
    6. What is the social utility of questions?
    7. Is test claimed to be diagnostic? (Is so, proof and see VI, 5, c, below).
  B. Statistical (10)
    1. Correlated against what outside criteria?
    2. Size of coefficient of correlation?
    3. Size and representativeness of sampling?
    4. Proof of validity of items (such as statements as to experimental

* A Scale for Rating Standardized Tests," *School of Education Record of the University of North Dakota*, Vol. 14, No. 1. Grand Forks: University of North Dakota, 1928, p. 11–15.

tryout of items individually to determine that no large percentage is failed or passed by all pupils and that the items show a consistent increase of percentages of successes with successive age or grade levels).

III. Reliability (25)
- A. Most important items.
  1. Correlated with what?
  2. Size and representativeness of sampling?
  3. Reliability coefficient.
  4. The means of the distributions.
  5. The standard deviations of the distributions.
  6. If some other measure than the above three is given to prove reliability, what is it?
  7. Intercorrelations.
- B. Less important but desirable.
  1. Order of giving various forms of test.
  2. Is test reliable enough statistically for individual measurement, or can it be used only for groups?
  3. Evenness of scaling (see II, B, 4).
  4. Are pupils accustomed to this type of test?

IV. Ease of Administration (15)
1. Manual of Directions (3)
   - a. How complete and simple is the manual?
   - b. Does manual control test conditions well?
   - c. Typographic makeup.
2. Simplicity of administration (8)
   - a. Amount of explanation needed for pupils by examiner?
   - b. Are directions to pupils clear, detailed, comprehensive?
   - c. Is arrangement of test convenient for pupils?
   - d. Are samples and "fore-exercises" given when needed?
3. Alternate forms (3)
   - a. Number.
   - b. Evidence of reliability.
   - c. Evidence of equivalency.
4. Time needed for giving.

V. Ease of Scoring (10)
1. Degree of objectivity—purely objective or some judgment on part of examiner?
2. Are adequate directions given—clear, equal to all emergencies?
3. Is scoring key adjusted to size of test?
4. Time needed to score one test.
5. Simplicity of procedure.
   - a. Number of processes needed to get final score?

VI. Ease of Interpretation (20)
1. Norms (6)
   - a. Kind—age, grade, percentile, etc.
   - b. Derivation—size and representativeness of sampling.
   - c. Tentative, arbitrary, or experimental?
   - d. For separate parts?
   - e. How expressed?
2. Is class record provided?
3. Are there provisions for graphing results?
4. Is interpretation of raw scores easy or hard?
5. Application of results (10)

a. Are directions or suggestions given for application of results to benefit teaching or administration?
b. Are tests survey or diagnostic?
c. If diagnostic—
(1) Proof of diagnostic value?
(2) What principle or principles underlie construction?
(3) How many different skills, abilities, or aspects of the subject are analyzed or measured?
(4) Does the analysis of total subjects into unit abilities follow teaching practices of needs?
(5) Is the diagnosis individual or class—proof?
(6) Does the test demand tabulations of individual pupils' errors to secure a diagnosis?
(7) Is a remedial program provided or suggested?

VII. Miscellaneous (5)
1. Typography and makeup.
a. Arrangement of printed matter.
b. Legibility of type.
c. Quality of paper.
d. Are test blanks free from distractions, norms, directions to examiner, etc.?
2. Is the time required for giving as small as is consistent with reliable measurement?
3. Is the cost in keeping with the amount, scope, and reliability of the results yielded?
4. Is good test service provided by the publisher?
5. Kind of new-type questions used?

points than that of Otis, but provides separate weights or scores only for the major divisions. Also a few points are mentioned for which no weights are given.

Such scales as these are undoubtedly helpful and the writer recommends that they be consulted by teachers at least until they become familiar with the criteria for selecting tests and their application. It should not be overlooked, however, that there is one possible fallacy connected with their use as with that of school building score cards and other similar instruments that sometimes may render the total score given by such a scale or score card a very unsatisfactory measure of the worth of the test or other thing being rated. For example, the Strayer-Engelhardt Score Card for City School Buildings, which weights different items on the basis of a total of one thousand points, assigns seventy of these points to heating and ventilation. It would, therefore, be theoretically possible for a building to receive a score of nine hundred thirty points, which is higher than practically any building ever rates, and yet to have absolutely no heat or ventilation and thus be absolutely unsuitable for use as a school building.

Such extreme cases, of course, rarely if ever occur, but it does happen occasionally that a building is so deficient in some one important point, although ranking fairly high on most others, that even though it receives what is considered a creditable rating upon the card, it is decidedly unfit for use. Similarly a test might conceivably be totally lacking in reliability and yet rate ninety according to Otis' suggested scale, or have no norms and yet rate ninety-five. Such unusual situations are, however, not very likely to occur. If they do the test in question should be ruled out of consideration, at least for most purposes, and no further attempt made to compute a complete rating for it.

The plan of describing tests in the following chapters.—In the following chapters the tests which the writer considers most deserving of use in the various school subjects, intelligence, and other fields of interest to teachers will be briefly described and criticized. In so doing the chief points discussed in this chapter will be considered, though not always mentioned.

The account of each test will begin with its exact title followed by the name of the author or authors and the date of copyright or first appearance if known. In many cases this date is not given upon the test. Since the writer does not consider the date of vital importance, he has not gone to much trouble to determine it in such cases, but instead has given the approximate date if he knows it or, if not, no date at all. Following this will be a statement of the parts into which the test is divided, the number of forms in existence, or, in the case of a series of tests, the number of tests in the series. Next the types of exercises, whether multiple-answer, true-false, completion, matching, or something else, and the exact or approximate number of elements in the test will be given.

In dealing with the content, information, if available, will be given as to the basis of selection, whether from courses of study, textbooks, examination questions, or other sources. Also if known something will be stated as to how thoroughly the test items were tried out and how much care was taken to eliminate unsatisfactory ones.

With regard to directions, nothing will be included if the directions may be regarded as satisfactory, unless they involve some unusual feature. If they appear to be unsatisfactory the point in which they seem deficient will be indicated. The exact

or approximate time limits will be given in each case. Likewise with regard to scoring directions and aids and objectivity of scores, nothing will be said if they may be regarded as satisfactory.

The function or apparent function of the tests will be stated unless it is well enough indicated by the title. Frequently some statement as to the validity of the test either on the basis of actual evidence or of inspection by the writer will be added.

Wherever possible reliability will be indicated by giving the four measures discussed in the section above under this topic, that is, the coefficient of reliability, the probable error of measurement, the ratio of this to the mean, and also to the standard deviation. These will ordinarily be inserted without discussion, except that occasionally information as to whether or not the number of individuals concerned in the determination of reliability was adequate will be added.

The best available norms will be included, although not always in complete form. The approximate number of cases or some other indication of how adequate the norms are will also be stated where known.

Following each test will be the name of the publisher without, however, the address. The latter may be found in Appendix A which contains a list of all publishers referred to and their addresses. Following the name of the publisher will be a statement of the cost, ordinarily first of a sample or specimen set and then of a package of the size in which the test is sold, or otherwise as stated in the publisher's catalog. Most publishers reserve the right to change prices without notice, but such changes are relatively infrequent, and since the writer has endeavored to get all prices at as late a date as possible, it is probable that very few will be changed within the next few years.

Finally under references the writer will give one or perhaps a few that describe most satisfactorily the test, its construction, derivation, and use. In many cases no such references are available, and in many others the best account of such points is contained in teachers' manuals, which are not listed among the references.

Summary.—The chief criteria which should be kept in mind in selecting tests are validity, reliability, objectivity, norms and

other provisions for using results, duplicate forms, scaling, ease of administration, and cost. Validity refers to whether or not a test accomplishes its purpose and may be subdivided into curricular and statistical validity. The former refers to agreement with the content of a desirable curriculum and the latter to the statistical testing of scores to determine their validity. Reliability is synonymous with accuracy and is determined by giving a test twice to the same pupils. Four measures of reliability, the coefficient of reliability, the standard or probable error of measurement, the ratio of this error to the mean, and its ratio to the standard deviation, are discussed. Among the chief factors that affect reliability are the length of a test, the scaling of the test elements, and the directions for giving and scoring. Errors may be classified as constant and variable, the former tending to be the same for all members of a given group, whereas the latter vary for different individuals. Objectivity refers to the quality of a test that there is no doubt as to what the correct answers are. Following this norms are discussed. The meaning of duplicate and equivalent forms is stated and methods of securing such forms given. The scaling of test elements is explained, although this procedure is not of vital importance in most cases. It is recommended that equal weight ordinarily be given the various elements of a test in scoring. Though not of great theoretical importance, the ease of administering a test and its cost are of practical importance. Two test rating scales, a fairly brief one by Otis and a longer one by Cole and von Borgersrode, are given. The chapter closes with a brief outline of the plan of describing tests to be following in succeeding chapters.

References

Monroe, W. S. "How to Make a Critical Study of an Educational Test," *An Introduction to the Theory of Educational Measurements.* Boston: Houghton Mifflin Company, 1923, Chapter IX.

Ruch, G. M. "The Criteria of a Good Test or Examination," *The Objective or New-Type Examination.* Chicago: Scott, Foresman and Company, 1929, Chapter II.

Ruch, G. M. and Stoddard, G. D. "Criteria for the Selection of Educational Tests," *Tests and Measurements in High School Instruction.* Yonkers: World Book Company, 1927, Chapter IV.

Smith, H. L. and Wright, W. W. "Criteria for Judging Standardized Tests,"

Tests and Measurements. New York: Silver, Burdett and Company, 1928, Chapter III.

Symonds, P. M. "Criteria for the Choice of Tests," *Measurement in Secondary Education.* New York: The Macmillan Company, 1927, Chapter XIV.

CHAPTER IV

ENGLISH AND RELATED SUBJECTS

Introduction.—Under this general head there will be included tests and scales in the fields of literature, language and grammar, composition, reading and vocabulary, spelling, speech, and writing. Of these subjects reading, spelling and writing often receive only very incidental attention in high school even in connection with the work in English, and the same is true of speech, except when there is a course specifically dealing with that subject. The other divisions of the field listed, literature, language and grammar,[1] and composition, may be said to receive no attention elsewhere in high school than in connection with English, but frequently compose practically the entire course in this subject. The subjects referred to just above should, however, also form a considerable part of all high-school English courses.

In high school as well as in elementary school the subject of English is of outstanding importance. It is to be expected, therefore, that the number of tests in the various branches of this subject should be, as it is, considerably greater than in any other subject. Since dealing with tests in all these branches in a single chapter would make it much longer than any of the others devoted to tests in particular subjects or groups of subjects, two chapters have been devoted to them. The basis of division is that just indicated. In this, the first of the two chapters, literature, language and grammar, and composition will be treated and in the next the other divisions of the field referred to above. Even though this separation is made the reader should not think of the two groups as being in any way fundamentally different but

[1] In the case of grammar an exception may be made to this statement in the case of the foreign languages, especially first-year work therein, since English grammar usually receives more or less emphasis in such courses.

rather as uniting to compose a single unified and integral content for high-school English, using the term in its broadest sense.

I. Literature

Although literature is perhaps most commonly thought of as the essential feature of high-school English, especially when it is contrasted with elementary-school English or language, the development of tests therein has been much slower than in grammar, reading, spelling, writing, and other subjects. There have been at least two important causes for this condition, neither one of which has been unique to literature as contrasted with other high-school subjects. One of these causes is the fact that there is a much larger element of what is commonly termed appreciation, the development of attitudes, ideals and points of view, involved in the teaching of literature than in that of the other branches of English, Since it is much more difficult to test appreciation than mere information, considerably less progress has been made in the field of high-school literature than in many others. In the second place, even with regard to the informational content of high-school literature, there has been and still is much less agreement among different schools and teachers than in the case of most of the elementary-school subjects. As a result it is difficult to formulate exercises for standardized tests which include facts generally enough dealt with in high-school English courses that the tests are suitable for use in a majority or even a fairly large proportion of high schools. In spite of these causes, however, there has been considerable activity along the lines of test making in the field of literature within the past few years and a number of tests have been constructed and made available which merit use. It cannot, however, even yet be claimed that the status of measurement herein is nearly as satisfactory as it is in most of the elementary-school subjects or even in many of the other high-school subjects, including most of the other phases of English. As will be seen from the tests discussed, practically all of what has been done has to do with the factual or informational phase of literature and in only two or three cases has anything really worth while been done along the line of testing appreciation.

Literature Test

E. R. Barrett and Teresa M. Ryan (1926)

Form A, B, C, also another

Forms A, B, and C each consist of from one to five multiple-answer exercises on each of from thirty-five to forty-two poems, plays, novels, and essays rather commonly read in high-school English work. The other form is composed of from four to twenty true-false statements on each of twenty-two selections. Many, but not all, of the selections dealt with in the forms are the same. Apparently the time allowed is thirty-five minutes, which should be enough that speed is not important. The items of knowledge called for are such that pupils who have not read the selections dealt with have little chance of knowing the correct answers. Forms A, B, and C yield very general measures of knowledge of literature and not satisfactory ones of knowledge of any of the selections dealt with. The other form, however, has enough elements on most of the selections that this is not liable to occur when it is used. Norms for May 1 are given as follows:

Grade	*IX*	*X*	*XI*	*XII*
Third quartile	20	22	27	31
Median	26	29	35	40
First quartile	34	38	45	51

Bureau of Educational Measurements and Standards. Sample set 20¢; 3¢ each, 65¢ per 25, $2.00 per 100.

English Literature Test

K. T. Omwake, R. E. Schwarz, and M. M. Ronning (1927)

Form 1

This test, which is one of the George Washington University Series, contains three subtests in multiple-answer, true-false, and matching form, with a total of 160 elements. English and American literature are chiefly dealt with, although there are a few references to that of other nations. Apparently the test is better suited for college or university than for secondary use, but it has been rather widely employed in high school, and has seemingly given at least fair satisfaction.

$r = .93$, $\text{P.E.}_{\text{meas.}} = 4$, $\frac{\text{P.E.}_{\text{meas.}}}{M} = .06$, $\frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .18$.

A correlation of .55 with school marks in English is reported. The working time is thirty-five minutes. The norms reported below are based on more than a thousand high-school seniors and more than five hundred in each of the other groups.

	Percentile				
	10	*25*	*50*	*75*	*90*
College students	58	79	98	117	129
High-school seniors	42	60	78	98	115
High-school juniors	23	33	48	62	79
High-school sophomores	18	28	38	49	63
High-school freshmen	13	20	30	43	55

Center for Psychological Service. 5¢ per copy, $4.50 per 100.

Objective Tests in Literature

S. R. Hadsell and G. C. Wells (1926)

These tests differ from the two previously described in that they test knowledge of each of a number of classics rather thoroughly instead of covering a large number more superficially. The series contains more than thirty tests dealing with as many books, poems, and plays commonly read in high school. The list includes Dickens' *A Tale of Two Cities*, Eliot's *Silas Marner*, Emerson's *Essays*, Goldsmith's *She Stoops to Conquer*, Lowell's *Vision of Sir Launfal*, several of Shakespeare's plays, and others. Each test consists of from three to five parts presenting a total of from sixty to one hundred fifty elements in true-false, multiple-answer, completion, matching, and direct-recall form. Considerable attention is given to the life and other writings of each author as well as to the work dealt with. The time limits for the longest of the tests are forty minutes, whereas for some of the others they are less. Tentative norms are announced for more than half of the tests.

Harlow Publishing Company. 10¢ per copy, $2.00 per 25; scoring key 10¢.

Objective Tests in English
Mabel S. Satterfield et al. (1926)

In both purpose and scope these tests are very similar to those of Hadsell and Wells. There are about fifty tests in the series, each covering a book or selection rather commonly read in high-school English. Most of the tests include about a hundred elements of several different types apiece, but several are much shorter. A few of the elements deal with the authors, but practically all are devoted to the works themselves. Time limits are not provided, but apparently each of the long tests requires at least half an hour.

Smith, Hammond and Company. Specimen set of 20 tests, $1.00; 5¢ per copy; $1.25 per 35, discount on large orders; shorter tests, 75¢ per 35.

Book Tests
M. W. Moe (1927)

There are four hundred of these tests covering books, poems, plays, and other works studied in high school. Each consists of a single five by eight card containing ten multiple-answer exercises upon one side. To indicate the scope of the list dealt with, the first ten alphabetically are as follows: *Abbe Pierre, The Able McLaughlins, About Paris, Abraham Lincoln* (Drinkwater), *Adam Bede, Adrift on an Ice Pan, Adventures in Contentment, Adventures in Friendship, Adventures of Captain Horn, African Game Trails.* No time limits are provided, but five minutes of working time should be ample for each. It is evident that tests calling for only ten items of knowledge are too short to yield anything approaching true measures of what pupils know about selections studied, but they are sufficient to indicate rather conclusively whether pupils have read and remembered the works covered or not. It appears, therefore, that these tests are much better suited for use in checking up on outside reading, which apparently is their chief purpose, than in connection with selections studied in class.[2] The method of scor-

[2] Although pupils may give their responses on the cards containing the exercises, apparently this is not intended. Answer slips containing space for ten responses are provided. It seems as if it might be intended that

ing is rather unusual, providing an answer code to be memorized by the teacher after which the printed copy is to be destroyed. The code is very simple, and the same one applies to all the tests.

Kenyon Press Publishing Company. 5¢ per copy; set containing one test on each of 100 titles $3.75, 2 sets $7.00, 3 sets $10.00, 4 sets $12.75; answer slips $1.00 per 1,000, $4.00 per 5,000, $7.50 per 10,000.

Literature Tests

H. T. Eaton (1928)

Tests 1–27

Each of these tests covers a different selection more or less commonly studied in high-school literature. The first five selections, which are as follows, will serve as examples of the whole list: "A Tale of Two Cities," "Franklin's Autobiography," "Burke's Speech," "MacBeth," "Julius Caesar." Each test contains three, four, or five parts dealing with character, setting, word study, plot, general background, identification of characters, and so forth. These parts are in multiple-answer, completion, and other forms, and in most cases include fifty elements. No time limits are stated, but apparently fifteen to twenty minutes should be enough on most of the tests.

Palmer Company. 2¢ per copy; 5¢ additional on each order for less than 15 copies.

Exercises in Judging Poetry

Allan Abbott and M. R. Trabue (1920)

Series X, Y

Each of the two series constitutes a scale for judging the quality of poetry and consists of thirteen sets each containing four versions of the same poem. One is the original; in another the

the teacher should read the test or copy it upon the board, although it may be that only a single pupil is supposed to be tested at a time. The answer slips are so simple, however, that they may very well be dispensed with, each pupil preparing his own by merely placing his name, the name of the book, and numbers from one to ten, on a blank sheet.

emotion has been falsified by introducing silly, affected, or otherwise insincere feelings; in another the original imagery has been made more commonplace and prosaic; and in another the metre has been so changed as to render the movement awkward, or at least less fine than in the original. Thus the qualities dealt with are emotional tone, imagery and rhythmic form. Almost all the originals are selected from writers of high excellence, including Scott, Tennyson, Shakespeare, Browning, Keats, and Burns. More than a hundred poems were tried out in selecting the twenty-six included in the two series. Pupils are allowed forty minutes to select the best and worst of each group of four. Each series begins with a simple four-line Mother Goose poem, and progresses to more difficult ones. The two series were submitted to over three thousand elementary and high-school pupils, college and university students, and others, and their responses tabulated. The results show that for many groups, especially those less mature, the original is not rated best by the largest per cent, either because it is too difficult to understand or because there is some definitely displeasing feature contained in it. The following norms in terms of the frequency of right choices of the "best," or original version, are given. They are based on two or three hundred in each group. In interpreting them it should be kept in mind that by mere chance one-fourth of the replies, or 3.25, would be correct.

	SERIES X									
	Elementary		*High school*				*College*			*Graduate students in English*
	VII	*VIII*	*I*	*II*	*III*	*IV*	*I*	*II*	*III-IV*	
Third quartile	5.1	5.2	5.7	6.3	6.5	7.5	8.5	9.0	9.9	11.6
Median	4.0	4.1	4.7	5.1	5.2	6.0	6.8	7.1	8.0	9.5
First quartile	2.8	3.1	3.7	4.0	4.1	4.5	5.3	5.4	6.2	7.4
	SERIES Y									
Third quartile	5.0	5.4	5.7	6.0	6.6	7.5	8.7	9.1	10.1	11.7
Median	4.1	4.3	4.6	4.9	5.4	5.9	6.9	7.4	8.0	9.6
First Quartile	3.1	3.3	3.4	3.9	4.0	4.3	5.3	5.5	6.5	7.7

Correlations between scores on Series X and Y do not indicate high reliability. Indeed, for elementary-school pupils there is practically no agreement. For high-school pupils the results are:

$$r = .44,\ \text{P.E.}_{\text{meas.}} = 1,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .16,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .50.$$

For college students they are somewhat more reliable.

Bureau of Publications. Specimen set 35¢; 5¢ per copy; manual 25¢.

Reference: Abbott, Allan and Trabue, M. R. "A Measure of Ability to Judge Poetry," *Teachers College Record*, 22:101–26, March, 1921.

Tests for the Appreciation of Literature

Hannah Logasa and Martha McCoy-Wright (1926)

The six tests of this series deal, respectively, with the discovery of theme, reader participation, reaction to sensory images, comparisons, trite and fresh expressions, and rhythm. Another on standard of taste in poetry was in the preliminary series but was later dropped. Test I presents ten poetic selections for each of which a one-word theme is to be given. In Test II are twelve selections, all poetry except two, for which pupils are to indicate the emotions aroused. Test III contains five poetic and one prose selection, for each of which the types of sensory images conveyed are to be indicated. Test IV presents ten very brief selections of poetry involving comparison and requires that each be marked as true, far fetched, or mixed. The fifth directs that all trite and fresh expressions in seven selections of poetry and one of prose be indicated. The last test contains ten selections, all poetry except one, and ten sets of straight or curved lines suggesting the swing of the various phrases. The proper connections are to be made. About fifteen minutes are necessary for each test.

This is the only series of tests for the purpose of measuring literary appreciation that may be considered at all standardized and this probably should still be regarded as largely experimental. The six phases of literary appreciation dealt with do not appear to compose the sum total of such appreciation. Further-

more to test appreciation of literature in general tests should not deal so much more with poetry than with prose. Tentative norms are given as follows:

Test	*High school*				*College*			
	I	*II*	*III*	*IV*	*I*	*II*	*III*	*IV*
I	5.0	5.3	5.6	5.9	6.5	7.5	8.5	8.8
II	4.8	5.0	5.2	5.4	5.8	6.0	6.3	6.8
III	6.4	6.5	6.6	6.7	6.7	6.8	6.9	7.0
IV	4.4	4.6	5.0	5.2	5.5	6.0	6.5	6.7
V	4.7	5.2	6.6	7.6	8.0	8.4	8.7	8.9
VI	3.5	3.8	4.1	4.5	4.6	4.7	4.9	5.0
Total	29	30	33	35	37	39	42	43

Public School Publishing Company. Sample set 15¢; in quantities, 10¢ per set.

Test of Literary Vocabulary

Laura H. Kennon (1925)

Forms A, B

Although this test appears to be intended chiefly for college and university use, it seems to merit some use in the last two years of high school also. Each form consists of one hundred words with five suggested meanings. The words were chosen largely from special fields of English because of their occurrence in supposedly familiar or famous passages of English prose or poetry, in the work of certain periods or authors included in the special field of English literature, in specific passages associated with special periods of history or with modes of thinking and living that have become a part of our literary background and general social inheritance, and in the so-called technical vocabulary of names of types of literature, figures of speech, critical terms, and so forth. They are arranged in cycles of seven according to their source. The coefficient of reliability of two preliminary forms which contained the same words as the final forms, but had them divided differently, was found to be .89 Also these preliminary forms correlated about .60 with the Army Alpha and Thorndike Intelligence Scales, about .50 with the North Carolina Exercises for Judging Prose, about .80 with final examinations in English, and almost .70 with semester marks in

English. It is stated that two other duplicate forms, C and D, are in course of preparation. A median score of 48 or 49 is reported for college graduates and teachers of English.

Bureau of Publications. Specimen set 20¢; $6.00 per 100.

References: Kennon, Laura H. V. "Tests of Literary Vocabulary for Teachers of English," *Teachers College, Columbia University, Contributions to Education*, No. 223. New York: Bureau of Publications, Teachers College, Columbia University, 1926. 78 p.

————. "Kennon Test of Literary Vocabulary," *English Journal*, 15: 61–63, January, 1926.

Reading Scales in English Literature

M. J. Van Wagenen (1921)

Forms A, B, C, Alpha, Beta

Since these scales are parts of a series which includes not only English literature but also other subjects, they are described in connection with the whole series on page 143.

Columbia Research Bureau English Test

H. R. Steeves, Allan Abbott, and B. D. Wood (1925)

Forms A, B

This test, although included here with those in literature, is broader in its scope, covering what the authors state are the four requisites in the study of English, spelling, mechanics including punctuation, vocabulary, and literary knowledge. Part I, on spelling, contains forty groups of four spellings of the same word, only one of which is correct. In the second part is a paragraph of about three hundred words containing a number of errors in grammar, syntax, punctuation, capitalization, use of idioms, and so forth, which are to be corrected. Part III, on vocabulary, contains a hundred words with four possible meanings of each, one of which is correct. The last part is composed of one hundred elements in multiple-response form, dealing with literature rather commonly studied. Most of them have to do with characters, a few with events, and a very few with authors. Almost two hours are required for this test. On the whole, it ap-

pears well suited for both high schools and colleges, as it is claimed to be. A few of the items, especially some of those in the subtest on literary knowledge, appear to be such as are not usually studied in high school, but rather in college, if at all. The test is practically self-administering, so that the person giving it needs to do little more than to distribute the booklets and call time. Largely because of its length, this test has a high reliability.

$$r = .97,\ \text{P.E.}_{\text{meas.}} = 4,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .03,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .12.$$

It correlates .60 with college entrance intelligence test marks and from .40 to .65 with final marks in college freshman English.

Norms for about eighteen hundred secondary school pupils are:

Year	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
XII	131	140	158	180	201	220	230
XI	118	128	146	167	188	208	222
X	101	112	128	151	172	189	202
IX	93	103	119	140	160	177	187

Those for the separate parts for 607 entering freshmen and 745 extension elementary English students at Columbia University are:

Part	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
I	29	30	34	36	37	39	40
II	20	23	26	32	38	43	48
III	35	41	50	62	74	83	88
IV	17	22	28	37	47	55	62
Total	113	125	146	167	194	216	228

World Book Company. Specimen set 30¢; $1.50 per 25.

II. Language and Grammar

Although some writers and test-makers use the two words in the heading with distinct meanings, others do not. Furthermore, several of the tests in this field contain exercises some of which belong under one heading and some under the other, if a distinction is made. It is, therefore, impracticable to separate these two phases of English for the purposes of this discussion. In general the term "language" is employed to refer to usage regardless of rules or principles, and "grammar" to knowledge of the latter.

Practically all the tests described in this section were either constructed primarily for use in the elementary school or for use there as well as in the secondary school. The few not of this type differ from the majority chiefly in being more difficult. Especially in the case of those constructed chiefly for elementary-school use, the selection of those to be included here was not easy, chiefly because of the large number of such tests now available. Also the fact that some which appear rather well suited for testing high-school pupils have received so little use for this purpose that no norms are available rendered selection more difficult. It is entirely possible that it might be desirable to use easier tests than most of those listed here if the entering class of a particular high school were very much below the average in their language ability. If this procedure seems desirable, the reader is advised to consult some one of the several books devoted chiefly to testing in the elementary school.

Grammar Test
T. J. Kirby (1920)
Forms 1, 2

This is in many ways similar to the well known Charters Diagnostic Language and Grammar Tests, but appears to be better suited to high-school use. There are five sections, each containing from eight to ten sentences and approximately the same number of grammatical rules or principles. In each sentence a choice must be made between two suggested forms, pronouns in the first two sections, verbs in the third and fourth, and miscellaneous usages in the fifth. After selecting the correct one of

the two forms pupils must choose from the rules or principles given the one which justifies the choice of forms. Thirty-five minutes are allowed for actual work. Provision is made for four scores for each pupil. These are, respectively, the number of sentences and of principles attempted and the number of each correct. The content of the test was determined by a study of a large number of errors actually made by pupils, from which the most common ones were chosen for inclusion. The scores yielded possess considerable diagnostic value. The reported measures of reliability for correct principles scores average about as follows:

$$r = .90,\ \text{P.E.}_{\text{meas.}} = 2,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .05,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .21.$$

Correct sentences scores are somewhat less reliable, as shown by the following figures;

$$r = .60,\ \text{P.E.}_{\text{meas.}} = 2,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .06,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .42.$$

The following median scores are announced:

	Grade					
	VII	*VIII*	*IX*	*X*	*XI*	*XII*
Sentences correct	31	33	33	36	36	35
Principles correct	23	28	26	33	31	34
Sentences attempted	37	37	39	40	40	41
Principles attempted	36	36	38	38	38	39

It will be seen that these norms show little increase from grade to grade. This may be interpreted as meaning that the test does not afford a satisfactory measure of class progress, but probably should rather be taken to mean that so little grammar is taught in high school that pupils do not improve much in knowledge thereof, whereas if it were actually taught the test might measure it fairly satisfactorily. Correlations of about .30 and .60 between this test and Briggs' English Form Test and of about .50 and .60 with Charters' Diagnostic Test have been reported.

Bureau of Educational Research and Service. $1.75 per 100.

Reference: Kirby, Thomas J. "A Grammar Test," *School and Society*, 11: 714–19, June 12, 1920.

English Form Test
T. H. Briggs (1921)
Forms Alpha, Beta

Although intended for use in Grades VII–IX, this test may well be used at least in the tenth also. Each form consists of twenty sentences ranging from a very few words up to several lines in length. They are arranged in five cycles of four each, within each of which will be found the seven errors dealt with by the test. Thus there are a total of thirty-five errors to be corrected. The seven errors or test elements were selected from a study of lists of minimum essentials in written English. They are initial capitals, terminal periods, terminal interrogation points, capitals for proper nouns or adjectives, run-on sentences, possessive apostrophes, and commas before "but" coördinating the parts of a compound sentence. Briggs states that an ideal test would require original writing by pupils involving these points, but that this, of course, could not be insured. Furthermore, dictation of sentences involving them would require too great an amount of time and care in correction, therefore the proof-reading form was adopted. A tryout of it along with dictation showed a correlation of .79 between the scores and that there were almost exactly one and one-half times as many errors in it as in the copying from dictation. This number varied greatly for the different incorrect items, however, ranging from about three-fourths to more than three times as many. The test requires twenty minutes to give. One experimenter has found a correlation of only about .30 between this test and the Kirby Grammar Test. Another has found a similar correlation of about .60. A coefficient of reliability between the two forms of .76 has been found.

The following mean scores expressed as per cents of errors and based on from one to more than four thousand pupils in each grade are given:

	Low VII	*High VII*	*Low VIII*	*High VIII*	*Middle IX*
Initial capital	3	3	2	2	1
Terminal period	4	3	2	2	2
Terminal interrogation point	20	17	12	12	9
Capital for proper noun or adjective	59	56	53	49	53
Run-on sentence	58	49	46	37	30
Possessive apostrophe	71	63	61	52	58
Comma before "but"	29	24	24	24	21

The corresponding scores, in terms of numbers of correct responses, are, respectively, 23, 24, 25, 26, and 26.

Bureau of Publications. Specimen set 20¢; 80¢ per 100, $7.50 per 1,000; scoring stencils 10¢.

References: Briggs, T. H. "An English Form Test," *Teachers College Record*, 22:1–11, January, 1921.
Certain, C. C. "The Briggs Form Test in Use," *English Journal*, 12:244–57, April, 1923.

Diagnostic Tests in English Composition (1923)

(a) Capitalization—S. L. Pressey and Mrs. E. V. Bowers
(b) Punctuation—S. L. Pressey and Helen Ruhlen
(c) Grammar—F. R. Conkling and S. L. Pressey
(d) Sentence Structure—F. R. Conkling and S. L. Pressey

Forms 1, 2, 3, 4 of each

These tests are intended for use throughout Grades VII–XII. Each consists of about thirty exercises. In those on capitalization and punctuation pupils are to insert the proper capitals and marks of punctuation, respectively, in the given sentences, which in the first case contain no capitals except at their beginnings, and in the second no punctuation except periods at the ends. The tests on grammar and sentence structure present groups of four statements of which three are correct. The four tests may all be given within one period. These tests are similar to Briggs' in testing proof-reading ability. They are more diagnostic than most other tests in this field and appear to be among the few most helpful ones. They have received unusually wide use since Forms 1 and 2 were employed in the 1924–25 nation-

wide testing program of their publishers and 3 and 4 in the similar 1927–28 program.

Coefficients of reliability of .83 for the Grammar Test and of about .60 for each of the others are reported. The corresponding probable errors of measurement are all about one point, their ratios to the mean about .05, .08, .05, and .06 for the four tests in order, and to the standard deviation about .40 for all except the one in grammar for which it is less than .30. Intercorrelations of from .40 to .66 have been found among the various tests of this series. Median scores are announced as follows:

	Grade						*College freshmen*
	VII	*VIII*	*IX*	*X*	*XI*	*XII*	
Capitalization	17	17	18	20	20	21	23
Punctuation	10	12	13	15	17	19	20
Grammar	9	14	18	20	21	22	24
Sentence structure	9	12	14	15	16	17	19

The results of the last nation-wide use in October and November in which about forty thousand pupils took each test are, however, somewhat higher.

	Grade						*College students*
	VII	*VIII*	*IX*	*X*	*XI*	*XII*	
Capitalization:							
Third quartile	21	22	23	24	24	25	—
Median	18	20	21	21	22	23	22
First quartile	15	17	18	19	20	20	—
Punctuation:							
Third quartile	13	16	19	22	24	24	—
Median	10	12	14	17	20	21	24
First quartile	7	9	10	12	15	16	—
Grammar:							
Third quartile	17	19	21	24	26	27	—
Median	14	15	17	20	22	23	26
First quartile	11	12	13	15	18	19	—
Sentence structure:							
Third quartile	15	16	17	19	20	21	—
Median	12	14	15	17	18	19	18
First quartile	9	11	13	14	16	16	—

Public School Publishing Company. Sample set 15¢; capitalization or punctuation test 75¢ per 100; grammar or sentence structure test $1.50 per 100.

Reference: "English Composition," *Report of the Fourth Annual Nation-Wide Testing Survey*, Project No. II. Bloomington, Illinois: Public School Publishing Company, 1927–28. 32 p.

English Test
E. A. Cross (1923)
Forms A, B, C [3]

This test is fairly difficult, being intended for high schools and colleges. Each form consists of eight parts which deal, respectively, with spelling, pronunciation, recognizing a sentence, punctuation, verb forms, pronoun forms, idiomatic expressions, and miscellaneous faulty expressions. In all except that on punctuation some form of alternative response is called for. In that the proper punctuation is to be inserted. The total number of elements is 172 and the time allowance forty-five minutes. The test is based upon six years of work with preliminary forms in testing college freshmen. It does not appear to have been satisfactorily validated, however, and its reliability is not high.

$$r = .70,\ \text{P.E.}_{\text{meas.}} = 6,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .04,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .45.$$

Percentile norms for seven hundred college freshmen at time of entering are given as follows:

Percentile	*5*	*10*	*25*	*50*	*75*	*90*	*95*
Norm	120	125	132	140	149	155	159

As is evident from the norms just given the scores tend to bunch somewhat closely around the median. This appears to show that the test does not differentiate abilities very satisfactorily.

[3] Cross and R. C. Pooley are preparing another series of somewhat similar tests. There are to be three forms of these also. Each is to contain two parts, the first subdivided into six exercises and the second into four. The total number of elements is somewhat greater than in the Cross tests. These are not yet commercially available, but it is probable that they will be published by the World Book Company.

World Book Company. Specimen set 20¢; $1.20 per 25.

Reference: Harvey, Nathan A. "The Cross English Test," *American Schoolmaster*, 18:85–86, February, 1925.

Language Error Test

G. M. Wilson (1923)

Test 1, Forms A, B, C; Test 2, Forms D, E, F

The three forms of each test are printed together in a single booklet. Each form consists merely of a story of about three hundred words which contains twenty-eight common language errors. Pupils are to read the story and correct all mistakes therein by drawing lines through the wrong words and writing in the correct ones. The time allowed is not definitely specified, but it is suggested that in the upper grades not more than ten minutes will be required. The errors included were chosen in accordance with the results of studies of those made by pupils in four city school systems and in a number of consolidated schools. The most frequent errors were embodied in stories, the forms tried out, and finally the tests in their present form constructed. It is suggested that the three stories be used either in combination as a single test or separately at different times as three tests. Several studies of the tests yield data on reliability averaging about as follows:

$$r = .80,\ \text{P.E.}_{\text{meas.}} = 1.5,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .08,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .30.$$

May medians based on many thousand pupils are announced as follows:

Grade	*VII*	*VIII*	*IX*	*X*	*XI*	*XII*
Median	20	22	23	24	25	26

It will be seen that these medians show only slight increases from year to year during the high-school period. The purpose of the tests is announced as being primarily diagnostic, but evidently unless all three forms are used the number of errors concerned is too small to be of much diagnostic value for individual pupils. For a class as a whole the results do possess considerable value of this sort.

World Book Company. Specimen set 10¢; 80¢ per 25.

References: Wilson, G. M. "Language Error Tests," *Journal of Educational Psychology* 13:341–49, 430–37; September, October, 1922.
————. "Locating the Language Errors of Children," *Elementary School Journal*, 21:290–96, December, 1920.
————. "After-Test Value of Language Error Tests," *Second Yearbook of the Department of Elementary School Principals*. Washington, D. C.: National Education Association, July, 1923, p. 371–80.

English Minimum Essentials Test

J. C. Tressler

Forms A, B, C

Each form includes seven subtests which deal, respectively, with grammatical correctness, vocabulary, punctuation and capitalization, the sentence and its parts, sentence sense, inflection and accent, and spelling. Some of the exercises are in multiple-answer form, others in single-answer, others call for properly marking given sentences and the spelling for writing dictated words. The total number of elements is eighty-six. The test is intended to be diagnostic, but the various subtests are too short to give it much value for this purpose. Speed is not intended to be an element in determining the score. The suggested time allowances for slow pupils decrease from fifty-five minutes in Grade VII to forty in Grade XII. For a single grade,

$$r = .80,\ \text{P.E.}_{\text{meas.}} = 3,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .05,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .30.$$

Tentative norms based on over three thousand scores from several schools in three states are as follows for the end of the year:

	Grade				
	VIII	*IX*	*X*	*XI*	*XII*
Third quartile	51	59	65	69	73
Median	42	48	55	61	67
First quartile	35	41	46	53	59

Public School Publishing Company. Sample set 10¢; 75¢ per 25.

Diagnostic English Tests

Wakefield

Forms A, B

The fourteen parts deal with constructions, forms, especially of verbs, sentence formation, and so forth and consist of a total of 128 exercises, almost all of which are in multiple-answer form. The tests are fairly diagnostic of pupils' knowledge of grammatical terminology, but have little to do with functional grammar. In other words, they tend to deal with what was emphasized formerly rather than at present, at least by the best courses of study and texts. Nothing is suggested as to the amount of time to be allowed, but apparently about thirty-five minutes are sufficient.

Bureau of Administrative Research. Specimen set 20¢; 2½¢ per copy, $2.00 per 100.

English Essentials Test

Annie Ginsberg and Rewey B. Inglis (1927)

Tests for Grades VIII, IX, X; Forms Alpha, Beta, Gamma[4] of each

Each test consists of one part dealing with spelling, one with sentence recognition, and one with grammar. In the ninth- and tenth-grade tests punctuation is combined with sentence recognition, whereas in that for the eighth grade there is a separate part for punctuation and capitalization. The spelling involves writing sentences from dictation; punctuation and sentence recognition call for proper marking of material presented to the pupils; the grammar is mostly in completion form with two possibilities suggested for each blank. There are no time limits, but it is stated that about forty-five minutes are sufficient. The tenth-grade spelling words were those of the Minnesota State Syllabus for that grade most frequently misspelled. Apparently those for the other grades were selected in a similar fashion and the remainder of the material in the tests according to the judg-

[4] A new Gamma form appears each year.

ment of the authors based upon the course of study. April norms for first administration of the tests are given as follows:

Grade	*VII*	*VIII*	*IX*	*X*	*XI*	*XII*
Median	54	63	68	74	80	82

"Improvement medians" for ninth-grade pupils only for the three forms in order are given as 65, 77, and 82, the times of administration being September, January, and April.

University Printing Company. 1¢ per copy, 75¢ per 100.

Self Marking School Tests

NO. I—ENGLISH

F. L. Clapp and R. V. Young (1927)

Forms A, B

This test deals with capitalization, punctuation, word form, and grammar, and is intended for use in Grades V–XII. Part I contains fifteen multiple-answer exercises, each composed of a statement in which something is wrong or omitted and four possible places where the mistake may be. Part II is composed of sixty-nine completion statements in each of which the right one of two possibilities is to be indicated. Part III contains sixteen statements having to do with various grammatical facts and principles, for each of which an answer is to be selected from four suggestions. Twenty-five minutes are allowed for taking the test, enough that speed is scarcely a factor in determining the score.

The test has one decidedly unusual feature. As presented to the pupil, sheets are clipped together in pairs, so that he sees only the outside of each pair. By the use of carbon strips between the two sheets of each pair responses are transferred to the inside of the second sheet and fall alongside the correct answers. After the test has been given the clips are removed or the edges of the paper clipped so that pupils can see the correct answers alongside their responses, and thus correct their own errors. Apparently this test has not received much use, and it is mentioned here chiefly because of its unusual feature. There are other tests which provide for pupils scoring their own papers,

but this one by presenting the correct answers in close juxtaposition to the pupil's responses seems to insure greater correctness in so doing. The coefficient of reliability is .85. Median scores are given as follows:

Grade	*VII*	*VIII*	*IX*	*X*	*XI*	*XII*
Score	80	85	88	90	92	93

Age norms are also available.

Houghton Mifflin Company. $1.25 per 25.

Diagnostic Tests in Practical English Grammar

Evalin E. Pribble and J. R. McCrory (1928)

Tests I, II; Forms A, B, C of each

Test I is intended for Grades VII and VIII and Test II for high school and college. Each has seven parts dealing with the following topics: verbs, pronouns, adjectives and adverbs, nouns, miscellaneous constructions, sentence recognition, and redundancy. Test I calls for 127 responses and II for 174. Some are of the ordinary completion type, others of completion type with two suggested answers, others of alternative type, and others of completion type calling for the proper forms of, or substitutes for, given words. A time limit of forty minutes is set.

The publishers state that these tests are "by far the most comprehensive English tests yet published" and that "they possess high diagnostic value." They are based upon careful selection and trying out of items. It is intended that they be given at the beginning and end of each semester or term. Apparently they are to some extent planned for use in connection with Miss Pribble's text, *Correct English Usage*. The coefficient of reliability is .88 and that with final marks in a teachers college grammar course .69.

Tentative September medians for high school are as follows:

Grade	*IX*	*X*	*XI*	*XII*
Median	121	133	139	148

Lyons and Carnahan. 4¢ per copy, $1.40 per 25, $2.10 per 50.

PURDUE DIAGNOSTIC ENGLISH TEST

G. C. Brandenburg and J. M. Stalnaker (1928)

Form A

This contains eight subtests dealing with punctuation, grammar, choice of words, spelling, information, vocabulary, reading (description), and reading (information). Each of the first six consists of from fifteen to thirty exercises, mostly in multiple-answer form with from two to four suggested answers. Each of the two reading subtests presents a selection to be read, followed in the first case by twelve yes-no questions and in the second by eight multiple-answer exercises. In the spelling subtest are a mixture of correctly and incorrectly spelled words, the latter of which are to be spelled correctly. The total time required to administer the test is about forty-five minutes. The coefficient of reliability is given as .93 and the coefficient of correlation with marks in English as .65. Mid-year norms are:

	Test								
Grade	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Total*
IX	12	15	14	24	9	18	10	6	107
VIII	9	14	13	23	7	15	9	5	95
VII	8	13	11	23	6	12	8	5	87

Lafayette Printing Company. Specimen set, 15¢; $1.00 per 25.

TESTS IN ENGLISH

Sarah M. Mullen and Muriel Lanz (1928)

Preliminary Diagnostic Test in Grammatical Usage, Forms A, B; Diagnostic, Accomplishment Tests, Numbers 1, 2, 3, 4, 5, 6 of each; Final Accomplishment Test in Grammatical Usage, Forms A, B

Each form of the Preliminary Test consists of sixty multiple-answer exercises. Each of the Diagnostic and Accomplishment Tests has twenty sentences to be corrected or completed, sometimes by choosing the proper one of two suggested forms. The Final Test consists of 120 sentences each of which involves choice between a right and a wrong expression. The errors dealt with in

the tests were selected according to the results obtained from several studies of oral usage of junior and senior high-school pupils. As is implied by their subtitles, the first test is to be given at the beginning of a course, each of the Diagnostic Tests over a particular phase of the work, each of the Accomplishment Tests later over the same phase, and the Final Test at the end. The tests are accompanied by a series of exercises to be used after the Diagnostic Tests to correct the errors revealed thereby. No time limits are given. Apparently the Preliminary Test needs about fifteen minutes, the Diagnostic and Accomplishment Tests about five minutes each, and the Final Test about thirty minutes.

Ginn and Company. Complete set of tests and exercises to accompany them, 72¢.

English Diction Test
T. H. Schutte (1926)

This high-school and college test consists of a narrative and descriptive selection of about four thousand words in length. Four hundred ten words or short groups of words, of which more than half contain errors in diction, are enclosed in rectangles. Pupils are to check each rectangle in which there is an error. Directions specify that pupils be allowed all the time needed to complete the test, and it is rather difficult to say how much this should be since slow, conscientious pupils might spend a great deal of time thinking about some of the expressions. Apparently an ordinary high-school period would be decidedly insufficient for some pupils and perhaps for many. The errors included were observed in the spoken and written work of at least ten normal-school students after a record of such faulty diction had been kept for two years. The manual contains a glossary of errors in which the errors are explained and better expressions suggested. The test appears to possess considerable merit, but for high-school use one based upon errors made by high-school pupils would be better. The test is designed for use at the beginning of a semester to indicate the points needing attention. After it has been given it is intended that the papers

be returned to the pupils and each one provided with a manual so that he can study the points which he did not know. The test should then be repeated at the end of the semester or perhaps earlier.

Public School Publishing Company. Sample set without scoring key 15¢; 8¢ per copy; manual 10¢; scoring key 15¢; $3.00 per 25 tests and manuals.

IOWA PLACEMENT EXAMINATIONS, ENGLISH APTITUDE AND TRAINING

M. F. Carpenter, G. D. Stoddard, C. E. Seashore, and G. M. Ruch (1924)

Forms A, B

Since the whole series of Iowa Placement Examinations is described on page 380, nothing will be included here concerning these tests except the statement that they are probably among the best available.

III. English Composition

The methods of measuring English composition are decidedly different from those employed in connection with the phases of English already treated in this chapter. The usual method is to compare pupils' work with a series of samples or specimens of English composition ranging from very poor to very good instead of by administering tests of the ordinary sort. A few tests, as contrasted with scales, have been labelled tests of English composition, but they are rather tests of language and grammar and thus properly belong in the last section of this chapter.[5]

Some of the existing English composition scales are very general and represent no attempt to diagnose the elements of the pupil's performance nor even to distinguish between different

[5] The diagnostic tests of English composition by Pressey and others which were described in the last section of this chapter are examples of tests of this sort.

types of writing. Several, however, limit their content to one particular type of writing, most commonly narration, and a few go even further than this in that they deal with specific kinds of letters, for example. A limited number of scales are diagnostic with regard to certain phases of writing or pupil abilities which compose writing. Thus, for example, Van Wagenen's scales, which will be described later in this section, make provision for separate ratings on thought content, structure, and mechanics. Undoubtedly in composition as in most other subjects, the present tendency is to go further in this direction. It appears probable, therefore, that the small number of such scales now available will be increased in the near future.

Although a few scales of the first type referred to in the preceding paragraph, that is, very general ones, will be treated in this section, this will be done rather because of their historical interest than because it is recommended that they be employed in actual work. Instead, one should use a composition scale limited to at least one type of writing, such as narration, description, or exposition. In practically all cases in which there are scales available in subdivisions of these fields, they should be preferred to those covering one type as a whole. There are, however, so few of this sort that those of the other kind will probably have to be used most frequently for the present.

Teachers employing English composition scales should bear in mind that it is almost if not entirely impossible to make the scores assigned pupils' work as a result of their use as reliable and objective as those from many standardized tests, using the latter term in its narrow sense. The subjective element cannot be eliminated in the comparison of pupils' work with the specimens composing a scale. It has been shown, however, that careful study of a scale and of the principles of using it coupled with the right kind of practice will very largely reduce the unreliability and subjectivity found in the scores of teachers who are just beginning to employ such scales. At first there will probably be no gain in these respects over scores or marks given by teachers on what may be called the "general impression basis," but this condition should change rather quickly as teachers be-

come more expert in using scales. Hudelson,[6] for example, presents the amount of improvement in these respects for 157 judges who rated themes. After merely reading the scale used there was no decrease, but indeed a slight increase, in the variability of their marks. However, after they had studied, discussed, and used the scale for two hours the variability was reduced to about two-thirds of what it was at first, after four hours to about one-half, after six hours to about one-fifth, and finally after sixteen hours to less than one-sixth. In connection with these figures the reader should bear in mind that the increase in the agreement among the judges as to what scores should be given did not result from mere undirected practice with the scales. Indeed results have been reported by other investigators which seem to show that such practice alone is of only slight help in accomplishing the desired end.

One of the chief values of the use of composition scales lies in the development of standards in the minds of teachers. The writer does not recommend that teachers employ composition scales directly for rating every theme to which they assign a mark. He does, however, most emphatically believe that they should become familiar enough with the best of such scales that, partly as a result of actually using them and partly as a result of becoming familiar with them, they should have built up in their minds more definite and satisfactory standards of pupil achievement than they would have otherwise. After such standards have been developed they will not need to employ scales in all their rating in order to be able to score pupils' compositions with a greater degree of accuracy than if scales were in no way concerned in the matter.

In discussing the various scales mentioned in this section, little attention will be paid to reliability and norms. There appears to be so little difference in the reliability of ratings based on the several best scales that it affords little assistance in choosing one rather than another. In general, the same norms apply to the Hillegas Scale and the numerous others which use the same

[6] Hudelson, Earl. "The Effect of Objective Standards upon Composition Teachers' Judgments," *Journal of Educational Research*, 12:329–40, December, 1925.

score values, therefore they are not repeated with each, but are given in the description of Hudelson's scale.

Scale for Measurement of Quality in English Composition by Young People

M. B. Hillegas (1912)

This, the first standardized composition scale published, is today receiving very little use in actual school work, but deserves mention because of its historical importance and because of its relation to other scales now more commonly employed. It consists of ten specimens ranging from one of value zero to one of value 937.[1] The ten samples were selected from seven thousand as a result of ratings by over two hundred judges. The scale specimens of least merit are artificial products, those of medium merit were written by high-school pupils, and the best by college freshmen. The chief defects of the scale are that no two of the samples deal with the same topic, that they differ very greatly in length and type of writing, and that the intervals between them are rather uneven. Because of this and other less important objections, it was not long until various other scales more or less derived from this were constructed to take its place.

Formerly published by the Bureau of Publications.

References: Hillegas, M. B. "A Scale for the Measurement of Quality in English Composition for Young People," *Teachers College Record*, 13: 1–54, September, 1912.

Thorndike, E. L. "Notes on the Significance and Use of the Hillegas Scale for Measuring Quality of English Composition," *English Journal*, 2: 551–61, November, 1913.

Extension of the Hillegas Scale for the Measurement of Quality in English Composition by Young People

E. L. Thorndike (1914)

In this, the first extension of the original Hillegas scale, Thorndike included twenty-nine compositions arranged at fifteen degrees of quality, ranging from zero to ninety-five, in terms of

[1] Sometimes one decimal place, and more often two, are pointed off in the scale values.

Hillegas scale values divided by ten. Some of the original Hillegas specimens were retained, others were dropped. At the best and worst qualities represented there are only single specimens, but at those near the middle of the scale there are several. The extension is similar to the original scale in that the specimens differ greatly in length, topic dealt with, and form. On the whole it was, however, a considerable improvement over that of Hillegas, partly because it was somewhat more finely divided in degrees of merit, and partly because the inclusion of several samples at each of a number of steps rendered it more likely that one would be found somewhat similar to the specimen being rated. Its length, however, made it harder to learn and use. As later scales have appeared, it has also practically ceased to be used.

Bureau of Publications. 12¢ per copy, $10.00 per 100.

Nassau County Supplement to the Hillegas Scale for Measuring the Quality of English Compositions
M. R. Trabue (1916)

Another well known and, for some years, widely used scale, based upon the original Hillegas scale is that named above. It contained ten specimens, arranged at approximately equal intervals from zero to nine, inclusive, the values being equivalent to the original Hillegas values divided by one hundred. All except the best three specimens were written by elementary school pupils upon the same topic; "What I Should Like to Do Next Saturday." This scale is, therefore, much more unified than either the Hillegas scale or the Thorndike extension thereof. The length of the compositions is so short that they hardly represent satisfactory samples. On the whole, however, this was the most satisfactory scale in English composition for some years after it appeared, but is scarcely to be recommended now in comparison with the other more recent ones.

Bureau of Publications. 8¢ per copy; $5.00 per 100; manual 35¢.

References: Trabue, M. R. "Supplementing the Hillegas Scale," *Teachers College Record*, 18:51–84, January, 1917.

———. "Nassau County Supplement to the Hillegas Scale for Measuring the Quality of English Compositions," *Report of a Survey of Public Education in Nassau County, New York*, University of the State of New York Bulletin, No. 652. Albany: University of the State of New York, December, 1917, p. 160–63.

English Composition Scale
Earl Hudelson (1919)

This is another one of the group of scales rather closely connected with that of Hillegas. Approximately one thousand compositions written by Virginia high-school freshmen were rated by an experienced scorer according to the Nassau County Supplement. After confirming this rater's reliability,[8] one hundred of the compositions ranging from poorest to best were selected, reproduced and scored by ninety-six composition teachers. Eleven specimens that, according to their median ratings, came very close to the even steps and half steps from two to seven, inclusive, were chosen from the hundred. Specimens for steps 7.5, 8.0, 8.5, and 9.5, were taken from 150 compositions collected and arranged for experimental use by Thorndike,[9] and the specimen at step 9.0 from the Thorndike Extension. The samples secured from Virginia all deal with the same subject, "The Most Exciting Ride I Ever Had," whereas the others are on various subjects, some of which are more or less similar to the one just given. Hudelson's suggestions for securing samples of pupils' work are that they be given the title used in the Virginia survey and told to write the best story they can on that subject in thirty minutes. Since this scale was largely based upon the work of Hillegas, Thorndike, and Trabue, it is probably a better scale than any one of the three. In some ways it is an advantage to have finer steps, as this scale does, and the total number is perhaps not too great. On the other hand, the objection has been made that the steps are finer than necessary or desirable because they call for greater discrimination than can be made reliably, at least by persons not highly trained in its use. No

[8] It was found that his judgment averaged only one-seventh of a scale step from the average judgment of ten trained scorers.

[9] Thorndike, E. L. *English Composition: 150 Specimens Arranged for Use in Psychological and Educational Experiments*. New York: Bureau of Publications, Teachers College, Columbia University, 1916. 127 p.

specimens below step two are included, which perhaps is a slight defect, although in actual practice high-school pupils will almost never produce compositions poor enough to be rated that low.

Although this scale received considerable use for some time and is still available, it has been largely superseded by the Hudelson Typical Composition Ability Scale, which will be described later.

World Book Company. 25¢ per copy.

Reference: Inglis, A. J. (Director). *Virginia Public Schools. A Survey.* Yonkers, New York: World Book Company, 1920, Part II, p. 213–22.

Typical Composition Ability Scale
Earl Hudelson (1922)

This, another one of the many scales which has employed the same values as the Hillegas scale, represents the outcome of one of the most careful and comprehensive studies along this line. During the past few years it has probably been the most widely used scale in this subject. It consists of eighteen compositions possessing scale values of from zero up to nine,[10] dealing with a common subject, "A Snowball Fight on Slatter's Hill." There is only one specimen at each of the two or three lowest and highest values, whereas at each of the intermediate values there are two or three. Hudelson had 770 junior and senior high-school pupils write upon thirty-two different representative assignments. The one that produced responses most typical of the pupils' average quality of composition work was chosen. Fifty selected compositions were reproduced and rated by two hundred composition teachers who had been trained in using scales. For rating purposes they used the earlier Hudelson English Composition Scale. On the basis of their ratings those compositions which appeared most suitable, were selected for the scale.

A rather unusual procedure is to be followed in getting samples of pupils' work to be rated by the scales. The original of the story dealt with in the scales is read to the pupils, the

[10] The exact scale values are not even integers, but are so close thereto, in no case varying by more than .05, that for all practical purposes they may be taken as such.

title and proper names being written on the blackboard. After the reading has been finished, pupils are given fifteen minutes in which to write the story. Thus their compositions are not measures of original ability, but of the formal side of composition ability. It is suggested that teachers just beginning to use the scale give ratings only at the values of the specimens it contains, that is, at intervals of one, but that after they become more expert intermediate ratings may well be given. Hudelson gives January norms based on the compilation of results from a large number of pupils, probably several hundred thousand, from many school systems in about twenty different states. These norms involve a combination of results from the original Hillegas scale, the Thorndike extension, the Nassau County Supplement, and Hudelson's own scale. They are as follows:

Grade	*VII*	*VIII*	*IX*	*X*	*XI*	*XII*
Quality	4.7	5.3	5.5	5.9	6.3	6.7

The median scores actually given the almost eight hundred compositions used in Hudelson's study were as follows:

Grade	*VII*	*VIII*	*IX*	*X*	*XI*	*XII*
Quality	4.6	5.5	5.6	6.2	6.4	6.7

Public School Publishing Company. Sample set 20¢; 10¢ per copy, $1.00 per 25; teacher's handbook 10¢.

Reference: Hudelson, Earl. "English Composition, Its Aims, Methods, and Measurement," *Twenty-Second Yearbook of the National Society for the Study of Education*, Part I. Bloomington, Illinois: Public School Publishing Company, 1923, Chapters V, VI, and Appendix VII.

English Composition Scales for Measuring Business and Social Correspondence

E. E. Lewis (1921)

Scales for Order, Application, Narrative and Expository Social Letters, Simple Narrations

These scales first appeared in three forms, varying in the number of samples included. The longest forms included about

thirty each, the next about half as many, and the shortest, the ones now commonly used, eight, nine, or ten. The Hillegas scale values are used, in this case, however, being divided by ten rather than by one hundred. The range of merit of the samples included in the short and commonly used forms varies somewhat, the poorest samples being rated at from zero to thirty, and the best at from eighty to ninety-five. The scale intervals are irregular, ten being their most frequent width. From the specimens that compose the longest form of each scale, those in the shorter scales were selected so as to have fairly even intervals from one to another.

For each scale several hundred pupils in Grades V–XII in a number of different school systems wrote letters or compositions and by process of elimination and repeated judging those included in the scales were chosen. Before being judged these were typed to eliminate the factors of handwriting, neatness, and so forth, but not of spelling, grammar, and other mechanics of English. Samples of advertisements and of "help wanted" ads were given the pupils to serve as the basis of the order and application letters. The social letters were secured by merely telling pupils to write them. For the narration scale 8,654 compositions from fifty-four different school systems were collected. The samples finally chosen for this were rated by 175 judges, more than in the case of the others.

If it is desired to measure letters of the types dealt with by four of these scales, these are probably the best instruments available for so doing. The scale dealing with narration does not, however, appear to possess as high merit as others which may be employed for this purpose. Lewis gives no distinctive norms for his scales, but states that those compiled by Hudelson apply to them as well as to the Hillegas and several others based more or less upon it.

World Book Company. Booklet containing all five scales 25¢ per copy.

Reference: Lewis, E. E. *Scales for Measuring Special Types of English Composition.* Yonkers, New York: World Book Company, 1921. 144 p.

English Composition Scales
M. J. Van Wagenen (1923)
Exposition, Narration, Description Scales

Each scale consists of fourteen or fifteen specimens that have been rated according to the Thorndike Extension of the Hillegas Scale, the original Hillegas values being divided by ten. Each has, however, been given three ratings, one for total content, one for structure, and one for mechanics. The poorest specimen on each scale has a rating somewhere between zero and ten on each quality, and the best one of about one hundred. In constructing the scales Van Wagenen had several thousand themes written by elementary and secondary-school pupils and college students on the topics "How I Earned Some Money," "When Mother Was Away," and "It Was a Sight Worth Seeing," rated. A group of high-school English teachers then selected forty or fifty themes from each group and these, along with some specimens from the Nassau County Supplement, were arranged in order of merit for each of the three specific elements by about 150 persons. As a result of these ratings the specimens included in the scales were chosen. Van Wagenen suggests that if a general merit rating is desired it be obtained by multiplying the thought content rating by four, the structure rating by two, and the mechanics rating by one, adding and dividing by seven.

On the whole, these scales rank well in comparison with the others available. All of the specimens in each deal with a single topic. The number of specimens included appears to be just about the optimum and to provide a great enough range of quality to include high-school themes ranging all the way from the very worst to the very best. The scales are unique in providing for rating in terms of the three qualities dealt with. Because of this feature they have much greater value for diagnostic purposes than scales that give a single general merit rating. On the other hand, it requires more time and labor to employ them and they are, therefore, not to be recommended when the purpose is merely to get a general rating of a pupil's ability to write themes. Also one hundred judges trained in using several

different scales reported that they found those of Van Wagenen rather bewildering. To assist in using the scales in a diagnostic or analytical manner Van Wagenen suggests several points to be taken into consideration in rating for each one of the three qualities.

Reliability coefficients of .67 for general merit, .70 for structure, .55 for thought and .50 for mechanics have been reported. These are low, but are based on such a small number of cases as not to be very conclusive. Also correlations of around .70 with school marks have been found. Separate norms are not given for these scales, but it is stated that general merit ratings are directly comparable with those from the Hillegas and other scales.

World Book Company. Booklet containing all scales, selected specimens for practice purposes, and so forth, 25¢ per copy.

A Scale for the Judgment of Composition Quality Only
S. A. Leonard (1925)

This is still another of the numerous scales that have used the same values as the Hillegas. It contains one or two samples at each integral step from one to seven, and also at steps 6.5, 7.5, and 8.5. All deal with the general subject of "Doing Something Difficult but Worth While," although most have more specific titles. The specimens vary greatly in length, some of the poorer ones being only one or two lines long, whereas some of the better ones are twenty to twenty-five lines. Following each are brief statements of its good and bad qualities. As a preliminary to the construction of the scale Leonard selected from several hundred themes written by pupils in Grades IV–IX, forty-eight which seemed most suitable for his purposes and later added thirty-nine others seemingly needed to fill in gaps in merit. All errors in form were corrected before the themes were duplicated. They were then rated on the Nassau County Supplement by about two hundred university students, half of whom had taught. These judges had first been given practice in the use of rating scales until their average variation was reduced to less than one-half step. Specimens were chosen for the scale which came nearest

the exact scale steps and showed the least deviation between medians in the two ratings.

As indicated by its title, the chief distinguishing feature of this scale is that there are no errors in the mechanics of English included, and that attention is centered entirely upon what Leonard calls "composition quality," or what has sometimes been called "story value." When it is desired to measure this apart from form, Leonard's scale is probably one of the best to employ, at least in the first two years of high school. For use in the junior and senior years there should be one or two samples at higher scale steps. The scale would be improved if there were at least two specimens at each step instead of only one as is the case at about half of the steps. Leonard gives no independent norms but quotes instead Trabue's actual medians and also suggested standards as follows:

Grade	*VII*	*VIII*	*IX*	*X*	*XI*	*XII*
Norm	4.7	5.3	5.2	5.9	6.3	6.7
Standard	5.0	5.5	6.0	6.5	6.9	7.2

National Council of Teachers of English. 15¢ per copy.

Reference: Leonard, S. A. "Building a Scale of Purely Composition Quality," *English Journal*, 14:760–75, December, 1925.

National Scales for Measuring Compositions
H. R. Driggs and A. F. Mayhew (1927)

In addition to scales for the lower grades these authors have published a booklet containing three scales for Grades VII, VIII and IX, respectively. There are five qualities in each labelled A, B, C, D, and E, with three specimens of each quality. There is also a short statement of the characteristics of each quality. The general topic dealt with is "The Cost of Carelessness," although in many cases more specific subjects head the specimens. Thirty-five thousand compositions by junior high-school pupils were collected, rated by teacher-students with a tentative scale, and from the results those included in the scales were selected. Rather detailed directions for securing samples of pupils' work to rate by the scales are given. They provide that

the topic mentioned above be discussed orally five or ten minutes, after which pupils are to be allowed fifteen minutes for writing their stories with no help and no recopying. The following scores were made by pupils of twenty-three schools near the beginning of the year's work, in terms of per cents of pupils whose compositions received each of the five letter ratings:

Grade	*A*	*B*	*C*	*D*	*E*
IX	12	25	35	20	8
VIII	14	24	33	21	8
VII	8	22	31	27	12

These scales are probably not so valuable as some of the others. Five degrees of quality seem to be too few to permit as exact rating as can be made with a fairly high degree of reliability. Also, as can be seen by the norms given above, the number of pupils receiving *A* ratings is great enough that it seems there should be samples of greater merit in the scales so that the best of the pupils might receive still higher ratings, and the same for those receiving *E* ratings and samples of less merit.

University Publishing Company. 25¢ per copy.

Letter-Writing Test

F. L. Clark (1926)

This test appears too easy for use in high school provided pupils have had satisfactory elementary training in letter writing. On the other hand, if they have not, and perhaps in some cases if they have, it may be employed profitably. It includes three subtests. The first lists fourteen expressions commonly used in opening or closing business and social letters and requires pupils to group them in six different lists according to use for such purposes as salutations in business letters, complimentary closes in social letters, and so on. In Subtest II six parts of a letter are named, followed by two blank letter forms representing the different parts of a business and a social letter with each part numbered. Pupils are to place the name of the proper part after each number. In Subtest III there are sentences and parts of sentences in confused order, which when properly put to-

gether will form a business and a social letter. Pupils are to arrange them in order and copy them in proper form, supplying their own signatures and addresses. Pupils are told to use as much time as is really needed, not to hurry, but not to waste time. It does not appear that the ordinary high-school or upper-grade class should require an ordinary period.

$$r = .85,\ \text{P.E.}_{\text{meas.}} = 3,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .06,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .26.$$

Norms for over ten thousand cases are as follows:

Test	Grade					
	VII	*VIII*	*IX*	*X*	*XI*	*XII*
I						
Third quartile	24	25	26	26	27	27
Median	18	21	22	24	26	26
First quartile	13	15	16	18	21	23
II						
Third quartile	11	11	11	11	11	11
Median	10	11	11	11	11	11
First quartile	7	9	9	11	11	11
III						
Third quartile	19	26	29	30	34	36
Median	11	18	23	24	28	31
First quartile	6	10	15	18	23	25
Total						
Third quartile	50	60	63	65	70	71
Median	38	49	53	57	64	66
First quartile	27	36	43	47	55	58

Since the maximum score on Part II is eleven, this portion is evidently much too easy for high-school pupils.

Public School Publishing Company. Sample set 15¢; $3.00 per 100.

BIBLIOGRAPHY[11]

I. Literature and General

Crow, C. S. "Evaluation of English Literature in the High School," *Teachers College, Columbia University, Contributions to Education*, No. 141.

[11] The bibliographies at the ends of the chapters are not complete, but merely give helpful references.

New York: Bureau of Publications, Teachers College, Columbia University, 1924. 172 p.

Dolch, E. W. "The Measurement of High School English," *Journal of Educational Research*, 4:279–86, November, 1921.

Du Breuil, A. J. "True-False Test in Literature and Formal English," *Illinois Association of Teachers of English Bulletin*, 15:1–17, May, 1923.

Irion, T. W. H. "Comprehension Difficulties of Ninth Grade Students in the Study of Literature," *Teachers College, Columbia University, Contributions to Education*, No. 189. New York: Bureau of Publications, Teachers College, Columbia University, 1925. 116 p.

Melvin, A. G. "A True-False Test in English Literature," *English Journal*, 11:491–96, October, 1922.

Odell, C. W. "Tests on Home Reading," *High School Conference Proceedings*, 1928. Urbana: University of Illinois, 1929, p. 104–8, 135–47.

———. "Tests on Outside Reading," *English Journal*, 18:827–32, December, 1929.

Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction*. Yonkers: World Book Company, 1927, p. 130.

Ruhlen, H. V. "Experiment in Testing Appreciation," *English Journal*, 15:202–9, March, 1926.

Smith, H. L. and Wright, W. W. *Tests and Measurements*. New York: Silver, Burdett and Company, 1928, p. 160–61.

Symonds, P. M. *Measurement in Secondary Education*. New York: Macmillan Company, 1927, p. 84–85, 109.

Trabue, M. R. "A New English Prose Test," *High School Journal*, 6:188–90, 214–19, 231–33; November, December, 1923.

Wilson, G. M. and Hoke, K. J. *How to Measure*, Revised and Enlarged. New York: The Macmillan Company, 1928, p. 454–55, 470–73.

"Tests to be Used in Measuring Appreciation of Literature," *School Review*, 33:491–92, September, 1925.

II. Language and Grammar

Ashbaugh, E. J. "The Measurement of Language: What is Measured and its Significance," *Journal of Educational Research*, 4:32–39, June, 1921.

———. "Senior High-School English as Revealed by a Standardized Test," *Journal of Educational Research*, 13: 249–58, April, 1926.

Briggs, T. H. "A Dictionary Test," *Teachers College Record*, 24:355–65, September, 1923.

Certain, C. C. "The Briggs Form Test in Use," *English Journal*, 12:244–57, April, 1923.

Charters, W. W. "Constructing a Language and Grammar Scale," *Journal of Educational Research*, 1: 249–57, April, 1920.

———. "Diagnosis of Grammatical Errors," *Sixth Conference on Educational Measurements, Bulletin of the Extension Division, Indiana University*, Vol. 5, No. 1. Bloomington: Indiana University, 1919, p. 13–24.

———. "Diagnosis of Language Errors," *Sixth Conference on Edu-*

cational Measurements, Bulletin of the Extension Division, Indiana University, Vol. 5, No. 1. Bloomington: Indiana University, 1919, p. 6–12.

Du Breuil, A. J. "True-False Test in Literature and Formal English," *Illinois Association of Teachers of English Bulletin,* 15:1–17, May, 1923.

Irmina, Sister M. "A Study of Language and Grammar Tests," *Catholic University, Educational Research Bulletin,* Vol. 1, No. 8. Washington, D. C.: Catholic Education Press, 1926. 40 p.

Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction.* Yonkers: World Book Company, 1927, p. 97–108.

Smith, H. L. and Wright, W. W. *Tests and Measurements.* New York: Silver, Burdett and Company, 1928, p. 132–45, 149–58.

Starch, Daniel. *Educational Measurements.* New York: The Macmillan Company, 1916, p. 108–13.

———. "The Measurement of Achievement in English Grammar," *Journal of Educational Psychology,* 6:615–26, December, 1915.

Symonds, P. M. *Measurement in Secondary Education.* New York: The Macmillan Company, 1927, p. 96–106.

Trabue, M. R. "Completion Test Language Scales," *Teachers College, Columbia University, Contributions to Education,* No. 77. New York: Bureau of Publications, Teachers College, Columbia University, 1916. 119 p.

Wilson, G. M. and Hoke, K. J. *How to Measure,* Revised and Enlarged. New York: The Macmillan Company, 1928, p. 455–61, 468–69.

———. "The Measurement of Language," *How to Measure,* Revised and Enlarged. New York: The Macmillan Company, 1928, Chapter VI.

"Standardized Tests of Ability to Use Correct English," *Public Personnel Studies,* 6:241–50, December, 1928.

III. Composition

Abbott, Allan, et al. *Composition Standards.* New York: Bureau of Publications, Teachers College, Columbia University, 1927. 66 p.

Ballou, F. W. "Scales for the Measurement of English Compositions," *Harvard-Newton Bulletin,* No. 2. Cambridge, Massachusetts: Harvard University, 1914. 93 p.

Briggs, T. H. "English Composition Scales in Use," *Teachers College Record,* 23:423–52, November, 1922.

Certain, C. C. "Are Your Pupils up to Standard in Composition?" *English Journal,* 12: 365–77, June, 1923.

———. "Why Not Include Standard Tests in Your Teaching Program This Term?" *English Journal,* 12:463–80, September, 1923.

Chou, H. H. C. *The Measurement of Composition Ability.* New York: Bureau of Publications, Teachers College, Columbia University, 1923. 107 p.

Dolch, Jr., E. W. "More Accurate Use of Composition Scales," *English Journal,* 11:536–44, November, 1922.

Gainsburg, J. C. "Fundamental Issues in Evaluating Composition," *Pedagogical Seminary,* 31: 55–77, March, 1924.

Gordner, Ida. "The Purpose and Use of a Composition Scale," *High School Journal,* 6:7–8, January, 1923.

Guiler, W. S. "Diagnosing Student Short-comings in English Composition," *Journal of Educational Research*, 14:112–19, September, 1926.

Hudelson, Earl. "The Development and Comparative Values of Composition Scales," *High School Journal*, 6:9–11, January, 1923.

———. "Effect of Objective Standards upon Composition Teachers' Judgment," *Journal of Educational Research*, 12:329–40, December, 1925.

———. "English Composition: Its Aims, Methods, and Measurement," *Twenty-Second Yearbook of the National Society for the Study of Education*, Part I. Bloomington, Illinois: Public School Publishing Company, 1923. 173 p.

———. "Making a Local Composition Scale," *School Review*, 33: 601–9, October, 1925.

Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction*. Yonkers: World Book Company, 1927, p. 123–30.

Smith, H. L. and Wright, W. W. *Tests and Measurements*. New York: Silver, Burdett and Company, 1928, p. 145–49.

Stark, W. E. "Measurement of Eighth Grade Composition," *School and Society*, 2:208–16, August 7, 1915.

Symonds, P. M. *Measurement in Secondary Education*. New York: The Macmillan Company, 1927, p. 85–94.

Van Wagenen, M. J. "The Minnesota English Composition Scales: Their Derivation and Validity," *Educational Administration and Supervision*, 7:481–99, December, 1921.

Willing, M. H. "Individual Diagnosis in Written Composition," *Journal of Educational Research*, 13:77–89, February, 1926.

———. "Measurement of Written Composition in Grades IV to VIII," *English Journal*, 7:193–202, March, 1918.

———. "Valid Diagnosis in High School Composition," *Teachers College, Columbia University, Contributions to Education*, No. 230. New York: Bureau of Publications, Teachers College, Columbia University, 1926. 64 p.

Wilson, G. M. and Hoke, K. J. *How to Measure*, Revised and Enlarged. New York: The Macmillan Company, 1928, p. 473–78.

———. "The Measurement of English Composition," *How to Measure*, Revised and Enlarged. New York: The Macmillan Company, 1928, Chapter VII.

Wolfe, C. S. "Topeka Composition Scale," *Illinois Association of Teachers of English Bulletin*, 9:1–12, March, 1917.

CHAPTER V

ENGLISH AND RELATED SUBJECTS (CONTINUED)

Introduction.—As was stated at the beginning of the last chapter, this one will be devoted to tests in reading and vocabulary, spelling, speech, and writing. These are all generally thought of as elementary-school subjects, but they have a place, and that a rather important one, in the high school as well. The mechanics of reading and writing, the ability to spell correctly several thousand commonly used words, and the power to express oneself orally in a reasonably clear and forceful manner, are supposed to be acquired by pupils before they leave the elementary school, but as all experienced high-school teachers know, this is a supposition that is frequently not true in practice. There is, therefore, need for testing the abilities of high-school pupils, especially those just entering, along the different lines just mentioned. The chief purpose of such testing is generally to determine whether or not the pupils can attempt high-school work without suffering too great a disadvantage or without being too great a burden to the teacher. A second and scarcely if at all less important purpose is to diagnose the condition of those pupils who are weak along these lines and thus to aid in giving them the most helpful remedial treatment that can be devised.

If the place of these subjects in the secondary school is granted, as it almost universally is, it follows that they tend to fall within the province of the English teacher. Although we do not commonly think of them when high-school English is mentioned, it is there that they should receive the major emphasis. Teachers of all other subjects should, of course, cooperate and make use of their many opportunities for helping pupils along these lines, but they can hardly be expected to take the lead in this matter and to administer tests of reading, spelling, and other such abilities.

I. Reading and Vocabulary

There is no question that in the high school as well as in the elementary school reading is the most important of this group of subjects. In practically all high-school work reading is the most frequently employed method of acquiring knowledge by study and the pupil who cannot read with at least fair speed and comprehension is seriously handicapped. There are, however, two phases of reading ability, especially of comprehension ability, that should be differentiated. One, which may be termed general, is distinctly the business of the elementary school, and in so far as the high school finds it necessary to deal with this it is doing work that should have been done previously. The other phase, which may be called specific or technical, has to do chiefly with the vocabularies peculiar to the various high-school subjects. It is, therefore, chiefly the business of the high school and to a considerable extent of the teachers of the particular subjects concerned rather than of the English teacher. In order to render development of reading ability of the sort just referred to at all easy of accomplishment, pupils must have fairly good general reading ability.

It is implied by what has just been stated that a number of different varieties of reading tests are needed in order that the different phases of total reading ability be measured. The most general reading tests provide for only a single score, which is usually a composite of speed, difficulty, and correctness. Quite a number of reading tests, however, provide for two scores, one of speed or rate and the other of comprehension. Although there is a tendency for these two elements to be positively correlated, there are many individual exceptions. Many pupils who can read rather rapidly understand what they read poorly, whereas others, although they comprehend well, proceed at slow rates. Therefore if a pupil is discovered to be unable to read non-technical material as well as he should, the first steps should commonly be to find out which of these two, rate or comprehension, is the cause of his difficulty.

Comparatively little has been done by way of testing reading abilities in various kinds of subject matter. Most general reading tests are composed of prose selections with perhaps an oc-

casional one of poetry, and almost none deal exclusively or even chiefly with the latter. There are a very few tests which draw all their content from a single field or subject as, for example, from physics or history, and thus test specifically the ability of pupils to read the kind of material they will encounter in certain subjects. There are also a few vocabulary tests and lists for particular subjects. In connection with reading and vocabulary tests of this sort the reader should be warned against one interpretation of results which has sometimes been made. This is that if pupils make low scores on such tests these scores indicate that they should not be allowed to carry the subjects from which the materials composing the tests were selected. This conclusion is valid if it is limited to mean that pupils scoring low should not begin the study of such subjects immediately in the usual way, but it is not valid if taken to mean that they should never undertake it. The proper interpretation of low scores of this sort is that pupils should either not carry the subjects in question or that they should have some special preparation to aid them in reading the content of these subjects when they do begin them. In practice, considering the exigencies of the actual situation as it is in most schools, this special training may usually best be given by the teacher of the subject concerned at the very beginning of the course.

It has been a very common practice for the authors of reading tests to prepare a series of two or more tests of differing difficulty. Frequently there is one test for the lower grades, one for the upper grades, and one for the high school, or perhaps only two, one for the intermediate grades and one for the junior and senior high schools. Therefore a number of the tests described in this section are the most difficult ones of series which also contain one or more easier ones for use with younger children. Few reading tests have been prepared exclusively for high-school use.

Standardized Silent Reading Test III

W. S. Monroe (1918)

Forms 1, 2

This test is the third of a series of which the other two are intended for the lower and upper elementary grades, respec-

tively. It consists of twelve short paragraphs each followed by a single-answer or multiple-answer exercise dealing with the content of the paragraph. The time allowed, five minutes, is enough that many high-school pupils finish the test. Both rate and comprehension are measured. Probably the chief unfavorable criticism of this test is that it is too short to yield a very satisfactory measure of reading ability. On the other hand, since it is so short, both forms can be given in much less than half of a class period and the scores combined. The following figures for reliability are partly based upon actual data for this test and partly estimated from similar data for Tests I and II of the same series. The first line is for rate, the second for comprehension.

$$r = .80,\ \text{P.E.}_{\text{meas.}} = 11,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .08,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .30.$$

$$r = .70,\ \text{P.E.}_{\text{meas.}} = 1.5,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .10,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .37.$$

Norms are as follows for Form I:

Grade	*Middle of year*		*End of year*	
	Rate	*Comprehension*	*Rate*	*Comprehension*
XII	96	30	100	32
XI	90	27	94	29
X	85	25	87	26
IX	83	23	86	24

For Form II they are several points higher.

Public School Publishing Company. Sample set 6¢; $1.00 per 100.

References: Monroe, W. S. "Monroe's Standardized Silent Reading Tests," *Journal of Educational Psychology*, 9:303–12, June, 1918.
Stone, C. W. "Improving the Reading Ability of College Students," *Journal of Educational Method*, 2:8–23, September, 1922.

Reading Examination, Sigma 3
M. E. and Laura C. Haggerty (1920)
Forms A, B

This examination consists of three subtests dealing with vocabulary, sentence reading, and paragraph reading. In the first are fifty words with four suggested synonyms or brief definitions of each. In the second are forty questions to be answered yes or no. They deal with matters of common knowledge, but include a number of difficult words. The third subtest consists of seven paragraphs following each of which are a number of true-false and multiple-answer exercises dealing with its content. Within each subtest the elements increase in difficulty. The content of the test was selected after a study of seventh and eighth grade readers and history texts, and only what was found to have received general usage included. It is intended to measure reading ability all the way from the fifth grade up through high school. The total score is a measure of general reading ability, but the scores from the three subtests have some diagnostic value. The total working time is twenty-eight minutes in addition to which the directions including fore-exercises consume enough time that practically a whole high-school period is necessary.

There is some difference between the reliability of this test as given in the test manual and as found in certain experimental work. The following are average figures:

$$r = .85,\ \text{P.E.}_{\text{meas.}} = 6,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .11,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .25.$$

The test has been found to correlate above .60 with a composite of grade location, age, and teachers' estimates and also with a composite of seven intelligence test scores. The following norms are reported:

Grade	VII	VIII	IX	X	XI	XII
Norm	68	76	84	90	96	102

World Book Company. Specimen set 45¢; $1.10 per 25; manual 25¢.

Reference: Symonds, P. M. *Ability Standards for Standardized Achievement Tests in the High School.* New York: Bureau of Publications, Teachers College, Columbia University, 1927. 91 p.

Reading Scale for the Understanding of Sentences

E. L. Thorndike and W. A. McCall (1920)

Forms 1, 2, 3, 4, 5, 6, 7, 8, 9 C

This is one of the two or three most widely used tests in reading. Each form consists of about ten short paragraphs with three or more direct questions, based upon the content of the paragraph, following each. Pupils are allowed thirty minutes of actual working time. The test was devised for use in the elementary school, but is difficult enough that it measures the ability of many high-school pupils. The arrangement of exercises is scalar, and the time allowance generous, so that the test evidently measures comprehension or power and not rate. The character of the paragraphs varies greatly, as also does that of the questions. This results in scores that are composites of many types of reading ability and hence are very general measures. As is shown by the following data, based on the average results from a number of different experiments, this test is not as reliable as many others.

$$r = .65,\ \text{P.E.}_{\text{meas.}} = 1.5,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .06,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .40.$$

This scale is distinguished by the fact that it was in connection with it that the now well known and commonly used *T*-scale system was first applied. Scores are, therefore, commonly given in terms of *T*-scores rather than of number of exercises correct. For high school at the end of each year these are as follows:

Grade	*VII*	*VIII*	*IX*	*X*	*XI*	*XII*
T-score	58	61	62	64	65	68

The corresponding numbers of questions correct range from twenty-five up to twenty-eight. These figures show that there is very little increase in scores during the high-school period.

Bureau of Publications. Specimen set 10¢; $2.00 per 100, $18.00 per 1,000.

References: Garrison, S. C. and Robertson, M. S. "Reliability of the Thorndike-McCall Reading Scale," *Peabody Journal of Education*, 4: 162–64, November, 1926.

McCall, W. A. "Proposed Uniform Method of Scale Construction: With Application to a New Reading Scale," *Teachers College Record*, 22:31–51, January, 1921.

Unspeeded Reading Comprehension Test

J. C. Chapman (1925)

This test consists of a single sheet on one side of which are thirty-one paragraphs of one or two sentences each. In the second part of each paragraph is one word that spoils the meaning. Pupils are to cross out this word. The thirty minutes allowed are amply sufficient to carry out Chapman's purpose and make the test measure comprehension and not speed. Although the general idea of the test is good, there are several undesirable features. Directions to both teachers and pupils are printed in very small type. On the page above the directions is a considerable amount of material, chiefly norms, which would be much better placed upon a separate sheet for teachers only, as its presence where it is is confusing. Despite these undesirable qualities this test is included here because it is one of the few constructed to measure comprehension alone. The following are approximate norms:

Grade	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
XII	22	24	26	29	30	30	31
XI	19	20	22	25	28	30	30
X	16	17	19	22	25	28	29
IX	13	15	17	20	23	26	27
VIII	10	11	14	17	20	23	25
VII	8	9	12	15	18	21	22

J. B. Lippincott. $1.00 per 50.

Iowa High School Silent Reading Tests
A. N. Jorgensen and H. A. Greene (1927)
Form A

This test, or rather series of tests, is intended to measure a number of different abilities that tend to make up general reading ability. Its various parts deal with paragraph comprehension of material from social science, literature and science, word meaning in social science, science, mathematics and English, sentence meaning, sentence organization, paragraph organization including selection of central idea, outlining and general organization, use of an index including general use, selecting key words and alphabetizing, and rate. The tests cover fifteen pages and require a total of almost an hour of working time. There are two parts, which may be given in separate periods if desired. Most of the material employed in the test was chosen from texts and other sources commonly read by high-school pupils.

A somewhat unusual method of determining the validity of this test was employed, based largely on the assumption that one function of school training is the preparation of high-school pupils to make adequate use of the books commonly employed in high school. The validity of the test according to this criterion is high. A thorough study was made of the reliability of each of the parts of the test in each of Grades VII–XII. This study showed that most of the subtests have reliability measures within the following ranges:

$$r = .80\text{–}.95, \text{P.E.}_{\text{meas.}} = 1\text{–}2, \frac{\text{P.E.}_{\text{meas.}}}{M} = .03\text{–}.05, \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .15\text{–}.30.$$

Needless to say, the reliability of the whole test is decidedly high. The intercorrelations among the various parts of the subtests range from about .10 to .90, and among the various subtests themselves from about .30 to .90, the average of the latter being about .60. A correlation of .52 with the Thorndike-McCall Scale is reported.

Percentile norms for each part of each subtest are reported, but will not be given here. Instead merely the total comprehension score, which includes all except the rate, and the rate score will be given.

Grade	*Comprehension*					*Rate*				
	Percentile					*Percentile*				
	10	*25*	*50*	*75*	*90*	*10*	*25*	*50*	*75*	*90*
XII	90	132	181	227	238	188	210	243	284	317
XI	79	127	157	203	220	191	219	255	295	320
X	74	109	145	180	198	170	199	221	263	305
IX	58	92	127	161	183	168	195	226	275	296
VIII	42	73	101	130	146	155	180	211	251	289
VII	37	53	87	116	130	143	181	209	270	306

Bureau of Educational Research and Service. $5.00 per 100; manual 20¢.

Reference: Jorgensen, A. N. "Iowa Silent Reading Examinations," *University of Iowa Studies in Education*, Vol. 4, No. 3. Iowa City: University of Iowa, 1927. 76 p.

High School and College Reading Test

G. M. Whipple (1925)

Forms A, B

Form A consists of a six-page selection dealing with the League of Nations Assembly, and Form B of one of seven pages dealing with the question of Japanese exclusion. Embedded in the material are twenty questions and directions to be answered or carried out as they are encountered. They call for such responses as underlining, checking, and writing in words. Pupils are given ten minutes in which to complete the test. This test is intended to determine how rapidly students can read and comprehend material of the sort encountered in daily work. It appears rather hard for many high-school pupils, but otherwise well suited to its purpose. As would be expected, correlations of from .50 to .60 have been found between scores on this test and on group intelligence tests. Approximate reliability data are:

$$r = .90,\ \text{P.E.}_{\text{meas.}} = 7,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .09,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .21.$$

Norms based on almost three thousand high-school and over six thousand college and normal school students are as follows:

Grade	Percentile						
	5	10	25	50	75	90	95
College graduate	7	8	10	12	14	16	17
undergraduate	5	6	8	11	13	15	16
Normal school	3	4	6	8	11	13	14
High school XII	4	5	7	9	12	14	16
XI	3	4	6	8	11	13	15
X	3	4	5	7	9	12	13
IX	2	3	4	6	9	11	13

Public School Publishing Company. Sample set 15¢; $3.00 per 100.

Stanford Test of Comprehension of Literature

Mary C. Burch (1928)

Tests I, II, III; Forms A, B, of each

Test I deals with narration and description, II with character and emotion, and III with exposition. Each consists of ten or twelve paragraphs chosen from books or selections commonly read in junior and senior high-school literature, with four multiple-answer exercises based on each paragraph. Among the sources of the paragraphs are: Dickens' *Christmas Carol;* London's *Call of the Wild;* Scott's *Ivanhoe;* Hawthorne's *Twice Told Tales;* Burns' *Lyric Poems;* Shakespeare's *Julius Cæsar;* Washington's *Farewell Address;* Ruskin's *Essays;* and so forth. Pupils are allowed twenty minutes of working time.

The functions of the tests are stated to be to differentiate between the reading abilities of junior and senior high-school pupils, to measure growth in reading ability, and to determine what books or selections are suitable reading material in so far as difficulty is concerned for particular individuals or groups. For accomplishing the last of these purposes it appears that there is not enough material on each level of difficulty in the tests. The average reliability of the single tests is about as follows for Grades VII–XII:

$$r = .88,\ \text{P.E.}_{\text{meas.}} = 2,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .09,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .23.$$

For the three tests combined:

$$r = .95,\ \text{P.E.}_{\text{meas.}} = 3,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .05,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .14.$$

The norms given below are based on an average of about one hundred pupils to each grade.

Test	Grade					
	VII	*VIII*	*IX*	*X*	*XI*	*XII*
I	15	18	19	21	26	26
II	17	22	21	24	30	30
III	15	19	19	20	26	26
Total	46	58	60	67	84	84

Stanford University Press. 75¢ per 25, $2.50 per 100, $10.00 per 500; all three tests, $2.00 per 25, $7.00 per 100, $30.00 per 500.

Purdue Reading Test

H. H. Remmers and J. M. Stalnaker (1928)

Forms A, B

This test contains ten selections or subtests. In the first are forty yes-no questions dealing with matters of common knowledge, but employing a number of rather difficult words. Each of the other selections consists of a paragraph or more followed by ten[1] multiple-answer exercises, true-false statements, single-answer questions, or completion statements, dealing with the content of what has been read. The quoted material is taken from sociology, psychology, civics, agriculture, biology, algebra, and fiction. Forty minutes are allowed for the test. The reliability is approximately as follows:

$$r = .86,\ \text{P.E.}_{\text{meas.}} = 6,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .09,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .25.$$

Mean scores based on a total of twelve hundred cases are approximately as follows:

[1] In one case there are eighteen instead of ten.

Grade	*VII*	*VIII*	*IX*	*X*	*XI*	*XII*	*College freshmen*	*College sophomores*
Score	42	52	53	57	70	71	85	93

Lafayette Printing Company. Specimen set, 45¢; $1.25 per 25.

Précis Test

I. C. Poley (1927)

In this test are eight selections from material commonly read or suitable for work in such subjects as English, history, science, mathematics, and so forth and following each five brief summaries or précis of the selection. Pupils are instructed to select the one summary that best expresses the thought and to mark each of the other four as inadequate or wrong. Forty minutes' time is suggested as sufficient and apparently for most high-school students should be more than enough. This test is intended to have diagnostic value and undoubtedly does so, but to make it highly valuable there should be several selections of each type of content. The following reliability data are probably slightly too low:

$$r = .80,\ \text{P.E.}_{\text{meas.}} = 2,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .09,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .30.$$

Norms for a total of one thousand cases are:

Grade	*IX*	*X*	*XI*	*XII*
Third quartile	21	26	30	34
Median	16	·21	26	30
First quartile	11	17	22	27

Public School Publishing Company. Sample set 15¢; 75¢ per 25.

Reading Test for Senior High Schools and Colleges

M. J. Nelson and E. C. Denny (1929)

Forms A, B

The first part of this test consists of one hundred words with

five possible meanings for each. In the second there are nine groups of four multiple-answer exercises each dealing with paragraphs selected from various school subjects such as history, education, commercial law, civics, and so forth. No information is available as to the basis of selecting the content, but apparently the words in the vocabulary test as well as the material in the paragraph test were chosen from high-school and college subjects. The coefficient of reliability is .91. Both high-school and college norms are available.

Houghton, Mifflin Company. $1.65 per 25.

Reading Scales in English Literature, History, and General Science

M. J. Van Wagenen (1921)

Forms A, B of each, also C, Alpha, Beta of English Literature

This series of reading scales is intended to test whether or not pupils can read efficiently material of the sort commonly read in high-school courses in these three subjects. Each scale consists of about fifteen paragraphs averaging perhaps three hundred words in length, followed by several questions, or, in the case of Alpha and Beta literature scales, one or more multiple-answer exercises. The history, science, and literature scales A, B, and C are intended to measure ability to comprehend what is read, whereas the Alpha and Beta scales are designed to determine ability to interpret. The scales are intended to be both prognostic and diagnostic. They should ordinarily be given at the beginning of courses in these subjects. Pupils who make low scores are evidently not able to study the subjects dealt with except at a considerable disadvantage, and may either be barred from so doing or given special training in the effort to prepare them to do so on more nearly equal terms with others. The scoring system is similar to that in a number of other tests by the same author and others, and involves transmuting the point scores into a different form. It is stated that this is not a speed test, and pupils are instructed to use all the time necessary to read and reread if desired. On the other hand, teachers are instructed to collect the test booklets at the end of

a forty-five minute class period. The following reliability figures for less than a hundred cases are given.

$$r = .85,\ \text{P.E.}_{\text{meas.}} = 2.4,\ \frac{\text{P.E.}_{\text{meas}}}{M} = .03,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .26.$$

The following norms are given as tentative. They apply to all the scales.

Grade	*VIII*	*IX*	*X*	*XI*	*XII*
Third quartile	83	85	89	92	97
Median	77	80	84	87	90
First quartile	72	74	77	80	83

Public School Publishing Company. Sample set 20¢; $3.00 per 100.

Reference: Van Wagenen, M. J. "The Van Wagenen Reading Scales in History, General Science, and English Literature," *Journal of Educational Research*, 3:314–16, April, 1921.

Vocabulary Test

L. M. Terman and H. G. Childs (1911)

This test, which is the one employed as a part of the Stanford Revision of the Binet-Simon Intelligence Scale, consists of one hundred words selected by rule from a dictionary. Although the list has received its widest use in connection with the Stanford Revision, it has also been published separately and employed alone. On the whole it is generally considered the standard vocabulary test. The hundred words which it contains are not always arranged in the same manner. For use in the Stanford Revision they are divided into two equivalent lists of fifty words each and arranged in order from easy to difficult. Sometimes the whole hundred are arranged in this order or alphabetically in a single list, sometimes they are in two parts, each alphabetical. Also the manner of response and method of scoring are not uniform. In the Stanford Revision the child is asked to state the meaning of each word orally and whole, half, or no credit is given. In the various written forms he is asked to define the

words, to choose the proper one of several synonyms for each, to mark each as to whether he can define it exactly, explain it in a general way, has only an indefinite idea of its meaning, or does not know it at all. There is usually no prescribed time limit. When given orally ten to fifteen minutes are usually amply sufficient for an individual to define as many of the words as he can. In the various written forms the time varies according to the type of response asked from perhaps ten minutes up to thirty or forty.

Since the words were selected at equal intervals out of a dictionary, the score yielded is a measure of total vocabulary and not particularly of familiarity with that most frequently employed. Terman states that this test has a much higher value than any other one test in the Stanford Revision, and that for children of English-speaking parents it probably has a higher value than any three other tests therein. When the method of giving the test and scoring used in the Stanford Revision and in some of the written forms is used, some subjectivity is necessarily present. However, this element does not appear to be great enough to be a very serious objection to its use. The reliability of the list has been shown to be fairly high, although no figures of the ordinary sort are available. Terman gives the following norms as being reached by 60 to 65 per cent of individuals tested:

Age	*12*	*14*	*Average adult (16½)*	*Superior adult (19½)*
Norm	40	50	65	75

Included with Stanford Revision of the Binet-Simon Intelligence Scale, see page 399.

Whipple's arrangement, C. H. Stoelting Company. $1.50 per 100.

References: Terman, L. M. *The Measurement of Intelligence*. Boston: Houghton Mifflin Company, 1916, p. 224–31.

Terman, L. M. et al. *The Stanford Revision and Extension of the Binet-Simon Scale for Measuring Intelligence*. Baltimore: Warwick and York, 1917. 179 p.

Test of Word Knowledge
E. L. Thorndike (1921)
Forms A, B, C, D

Each form consists of one hundred words followed by five others of which one is an approximate synonym. These words were selected from Thorndike's ten thousand word list.[2] The first four words were selected from the first thousand in the list, the next eight from the second thousand, and so on down until finally the last twenty-four were chosen from words not included in the ten thousand at all. Apparently pupils are allowed as much time as necessary. Although these tests were intended primarily for elementary-school use, they have been shown to be sufficiently difficult that even high-school seniors fail on many of the words. On the whole, this must be rated as one of the two or three best vocabulary tests available.

Average data on reliability from two sources yield the following approximate figures:

$$r = .90,\ \text{P.E.}_{\text{meas.}} = 3,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .07,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .20.$$

It has also been shown that there is some difference in difficulty of the different forms. A is approximately three points easier than B and D, whereas C is approximately three points more difficult. Average results from the four forms yield approximately the following norms, not however in all cases based on pupils in the same portions of the country:

Grade	*VII*	*VIII*	*IX*	*XII*
Norm	49	56	64	64

Bureau of Publications. Specimen set 15¢; $1.50 per 100, $13.00 per 1,000.

Reference: Thorndike, E. L. and Symonds, P. M. "Difficulty, Reliability, and Grade Achievements in a Test of English Vocabulary," *Teachers College Record*, 24:438–45, November, 1923.

[2] Thorndike, E. L. *The Teacher's Word Book*. New York: Columbia University, 1921. 134 p.

Tests of English Vocabulary

Alexander Inglis (1923)

Forms A, B, C

Each form consists of one hundred and fifty words alphabetically arranged and used in short statements followed by five words, one of which is an approximate synonym. The tests are intended to determine "the extent to which the student has acquired the intelligent general reader's vocabulary," vocabulary being used to refer to passive or silent reading rather than active. All the words in the *New Modern English Dictionary* were examined and classified and these lists checked and revised by comparison with the *New International* and *Standard Dictionaries*, and with the results of all vocabulary studies known to Inglis. Thus the tests have as their basis a very thorough and comprehensive study. By selecting words occurring at regular numerical intervals, a group of approximately three thousand was chosen and from this group by the same method the words used in the three forms selected. It will be seen that a number of other duplicate forms can very easily be made. About thirty minutes should be enough time for the tests.

$$r = .90,\ \text{P.E.}_{\text{meas.}} = 2.6,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .05,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .20.$$

Medians from many thousand individuals are as follows:

Grade	*IX*	*X*	*XI*	*XII*	*College freshmen*	*College graduates*
Median	45	63	78	87	105	129

Ginn and Company. $1.00 per 30.

Reference: Inglis, Alexander. "A Vocabulary Test for High-School and College Students," *The English Leaflet*, Vol. 23, No. 197. Boston: New England Association of Teachers of English, 1923, p. 1–13.

English Vocabulary Tests for High School and College Students

W. T. Markham (1929)

Forms I, II

In each test are 125 words used in sentences or phrases, followed by five others of which some one is practically synonymous with the given word. A list of one thousand was obtained by selecting words at regular intervals from the *Winston Simplified Dictionary* and adding a few taken at random from *Webster's Small Dictionary,* and from students' manuscripts. From this list those used in each form were chosen. The author claims that the words so selected represent "a true sampling of the field covered by the general reading vocabulary of both high-school and college students." It is definitely stated that this is not a speed test, so apparently all the time needed is given. It should not, however, require more than an ordinary class period, and perhaps even less. A coefficient of reliability of .88 is reported. The announced medians for Form I are based on only about a thousand cases in all.

Grade	*IX*	*X*	*XI*	*XII*
Median	65	73	82	90.

It appears that Form II is from one to three words easier.

Public School Publishing Company. Sample set 10¢; $1.00 per 25.

Technical Vocabularies of the Public School Subjects

Luella C. Pressey (1923)

Although these vocabularies are in no sense tests, it seems in place to include them here because of their value as sources of test material. They are published in fifteen sections containing a total of nineteen lists as follows:

1. Grammar (English, German, Latin, French) and Composition
2. English and American Literature
3. Arithmetic

4. Geometry and Algebra
5. American History
6. General Science
7. Biology
8. Chemistry
9. Geography
10. Physics
11. Physiology and Hygiene
12. Home Economics
13. Manual Training (Woodwork and Elementary Metal Work)
14. Art
15. Music

Each list contains a total of from 42 up to 1,654 words, about one-half of them having one thousand or more. The words are arranged alphabetically. Three styles of type are used to distinguish words in so-called "essential" vocabularies, those in "accessory" vocabularies, and a few others which it seems should be included. The lists were made up by systematic tabulation of all technical or unusual words appearing in the most widely used textbooks of the subjects dealt with, supplemented by the ratings of teachers as to their importance. From eight to twenty-three texts were used in the different subjects and from twenty-seven to over one hundred teachers rated each list of words. The words checked by more than half the teachers formed the essential vocabularies referred to above, those checked by from one-fourth to one-half the accessory vocabularies, and those checked by less than one-fourth but appearing in more than one-third of the texts the third group.

Public School Publishing Company. 5¢ per copy; 40¢ to $1.50 per 35.

Reference: Pressey, L. C. "An Investigation of the Technical Vocabularies of the School Subjects," *Educational Research Bulletin (Ohio State University)*, 3:182–85, April 30, 1924.

II. Spelling

Although the number of spelling tests available for use in the elementary school is large and many of these may appropriately be given pupils just entering high school, those included in this section will be limited to a few that may be employed throughout the secondary-school period. For the purposes of making an

inventory of the spelling ability of entering pupils other tests than those described may well be used, but since such tests should properly be classed with those for the elementary grades, they will not be discussed here.

It is perhaps somewhat more true of spelling than of the other subjects dealt with in this chapter that it receives attention elsewhere than in English. Many teachers of other subjects expect their pupils to learn the spelling of technical and other terms that occur in their work and devote time and attention to this end. History teachers, for example, frequently require pupils to learn how to spell correctly the names of important characters and places mentioned; science teachers expect their pupils to learn the spelling of technical terms encountered, and so on in most subjects. The writer believes that this practice is to be commended and that all high-school teachers should include among their instructional aims the ability on the part of pupils to spell correctly all important or frequently used terms that occur in the content of their courses. If this were consistently carried out, the responsibility of the English teacher with regard to spelling would not be nearly so great as it is with regard to most of the other branches included under the heading of this and the last chapter. All that she should be expected to deal with in this line are the words peculiar to English itself, and those of common enough occurrence in various subjects that they do not belong particularly to any one, and even in the case of these her responsibility is not entire but shared with other teachers.

Extension of the Ayres Spelling Scale
B. R. Buckingham (1919)

To the original Ayres scale, which consisted of the thousand words found to be most common in a large amount of written correspondence, Buckingham added 505 words chosen because they were common to a number of spelling books.[3] Although

[3] The one thousand words of the Ayres scale were arranged in twenty-six groups or columns in each of which all the words are of approximately equal difficulty according to the spelling of a large number of children. Norms for each column for each grade from the second up are given on the scale.

some of the words occur at most of the steps on the original Ayres scale, most of them tend to be toward the upper or more difficult end. A few are in six additional more difficult steps. The difficulty of the words was determined in the same way as for those in the original list except that the ninth grade was included, which had not been done previously.

Public School Publishing Company. 14¢ per copy; three or more, 12¢ each.

Reference: Ayres, L. P. "A Measuring Scale for Ability in Spelling." *Russell Sage Foundation Bulletin* No. E 139. New York: Russell Sage Foundation, 1915. 59 p.

Sixteen Spelling Scales Standardized in Sentences for Secondary Schools

Earl Hudelson, F. L. Stetson, and Ella Woodyard (1920)

These scales, frequently referred to as the Seven-S Scales, are based upon a determination of the second and third thousand most frequently used words. This determination was made on the basis of the work of Eldridge[4] and Cook and O'Shea,[5] Ayres' thousand most common words,[6] and Jones' spelling "demons."[7] Proper names of persons and places, and certain other varieties of words were eliminated from the resulting list, and the two thousand highest ranking words remaining given to a large number of pupils to determine their difficulty. Finally four hundred of the two thousand words which seemed to represent fairly the various degrees of difficulty were arranged in tests along with four words common to all and others from the Ayres list, and were widely tried out. On the basis of these results sixteen twenty-word lists were constructed. Twelve of these lists are of equal difficulty, each word in each being equal in difficulty to

[4] Eldridge, R. C. *Six Thousand Common English Words.* Niagara Falls, New York: (Privately printed), 1911.

[5] Cook, W. A. and O'Shea, M. V. *The Child and His Spelling.* Indianapolis: Bobbs-Merrill, 1914. 282 p.

[6] Ayres, L. P. "A Measuring Scale for Ability in Spelling." *Russell Sage Foundation Bulletin* No. E 139. New York: Russell Sage Foundation, 1915. 59 p.

[7] Jones, W. F. *Concrete Investigation of the Material of English Spelling.* Vermillion: University of South Dakota, 1913. 27 p.

the corresponding word in each of the others. Likewise the other four are of equal difficulty, but of considerably greater difficulty than the first twelve. Each word is one-tenth sigma more difficult than the preceding word. The words were incorporated in sentences that are to be read to pupils who, however, write only the words. About five minutes time should be enough for the actual giving.

There seems no doubt that more time and care was given to constructing and standardizing these scales than is true for any other spelling scales or lists intended for high-school use. As would be expected, their reliability is high. It is recommended that to secure reliable individual scores two scales of forty words be employed. The following February norms in terms of per cents of correct spellings are given:

Grade	*VII*	*VIII*	*IX*	*X*	*XI*	*XII*
Scales I–XII	66	74	80	85	89	91
Scales XIII–XVI	35	45	54	61	67	72

Bureau of Publications, Teachers College. Bulletin containing all the scales and other material 40¢.

Reference: Hudelson, Earl, et al. "Sixteen Spelling Scales Standardized in Sentences for Secondary Schools," *Teachers College Record*, 21:337–91, September, 1920.

National Spelling Scales for Junior and Senior High Schools

J. J. Tipton (1926)

Tests 1, 2, 3 of each

Each of these tests consists of slightly more than fifty completion statements in each of which the word to be spelled has been omitted. The complete sentences are read by the tester and the pupils write in the omitted words. No time is suggested, the directions merely saying that the teacher shall read the sentences, pausing long enough for pupils to write each word. It seems, however, that not more than ten minutes should be necessary. The words in the junior high-school scale were selected from the five most widely used modern spelling books with a

few more difficult ones added. Norms are given in terms of spelling ages and months as follows:

Score	*1*	*11*	*21*	*31*	*41*	*51*
Senior high school	14–0	14–10	15–9	16–9	17–10	19–3
Junior high school	11–0	11–10	12–8	13–8	14–9	16–3

National Publishing Society. 2¢ per copy; directions 10¢; class record sheet 2¢.

National Spelling Lists for Junior and Senior High Schools

J. J. Tipton and W. R. Shaw (1928)

Lists 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23

The twelve lists mentioned above are intended for use in the different half grades from the first half of the seventh on through the second half of the twelfth, respectively. Each consists of a number of words, in most cases between two and three hundred, selected after a study of about a dozen of the outstanding vocabulary and spelling studies and most up-to-date texts. The words in each list are arranged alphabetically, and are stated to "comprise the minimum number that should be learned by pupils of the corresponding half year." Apparently they are to be spelled in ordinary column form.

National Publishing Society. 4¢ per copy; in quantities 3¢.

A Standard High School Spelling Scale

Revised and Enlarged Edition (1928)

E. P. Simmons and H. H. Bixler

Forms I, II, III, IV

This series consists of four forms of a high-school spelling test of one hundred words to be pronounced in sentences for pupils to spell, and sixty-four test lessons each consisting of forty words. The words included are those left from the Commonwealth Investigation list[8] of five thousand most commonly used words,

[8] Horn, Ernest. "A Basic Writing Vocabulary; 10,000 Words Most Commonly Used in Writing," *State University of Iowa Monographs in Education*, Series 1, No. 4. Iowa City: University of Iowa, 1926. 225 p.

after abbreviations, most plurals and past tenses, and words spelled by 90 per cent or more of high-school freshmen had been eliminated. The scale location of the words was determined by two hundred spellings of each.

Results from a limited number of eighth-grade pupils yield the following reliability data:

$$r = .93,\ \text{P.E.}_{\text{meas.}} = 8,\quad \frac{\text{P.E.}_{\text{meas.}}}{M} = .12,\quad \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .18$$

Norms for the hundred-word test are given on the basis of about five thousand freshmen and three thousand senior pupils from two hundred and fifty communities in five states as follows:

Grade	*VII*	*VIII*	*IX*	*XII*
Norm	63	67	70	84

For each of the sixty-four test lessons also norms are given. These lessons are arranged in order of increasing difficulty. Likewise an alphabetical list of all the words included in the sixty-four lessons is given with the percentile norms for freshmen and seniors after each. The test given at the beginning was prepared in coöperation with the English departments of high schools of Atlanta, Georgia, and is designed especially to select pupils who may be excused from carrying spelling in high school. A score of eighty has been set as high enough to permit exemption.

Smith, Hammond and Company. Single copy, 44¢; in quantities, 33¢.

III. Speech

Little has been done by way of measuring the ability of pupils in oral speech, especially at the high-school level. A very limited number of pronunciation and oral-reading tests are available for use in the elementary school, but, unless the writer is mistaken, none of these tests are suitable for high-school use, and no norms are available on that level. Several attempts have been made to

construct measuring instruments of one sort or another for speech or public speaking, as it is sometimes called. It has been suggested that such measurement might well be made by means of a set of records containing samples of speech of various degrees of merit. Another suggestion has been that a rating scale be employed more or less similar to the scales used for rating teachers, pupils, and so on. In so far as the writer knows, however, only one attempt at measurement along this line has resulted in tests that may with any justice at all be called standardized, and for reasons that will appear in the description of this series it is very doubtful if their use will ever become at all common.

Speech Measurements

Smiley Blanton, Margaret G. Blanton, and Sara M. Stinchfield (1926)

This is an elaborate series of tests for the measurement of many phases of ability in speech. They are arranged in ten groups, one for preschool and kindergarten use, one for each of the eight elementary grades, and one for adult use. Some of the tests are the same for two or more grades, whereas others differ. Part I of the series consists of a subjective rating by the examiner based on ten different qualities of general behavior and speech, whereas Part II is objective in its nature. The different phases measured in Part II are articulation, response to pictures, oral and silent reading, including vocabulary, and lung capacity. A number of the tests employed are previously standardized tests such as the Terman-Childs Vocabulary, the Starch Reading, and so forth. It is stated that ordinarily all the tests for any one grade can be given in about twenty minutes, although more time may be required in the case of marked deviations from good speech. It is, of course, evident from the list of abilities tested that some of the tests must be given individually, although others may be given to groups. Provision is made for combining scores on all the objective tests into a single index, which is supposed to approximate the subjective rating given.

The following norms, apparently based on rather few cases, are given for certain groups:

	Articulation		*Oral reading rate*	*Silent reading rate*	*Spontaneous speech rate*	*Per cent of relevant words*	*Vocabulary*
	Test A	*Test B*					
UNIVERSITY STUDENTS							
Unselected	97.5	97.5	276	204–58	120–50	96	73.5
Speech Cases	93	94	186	204–58	120–50	93	72
ELEMENTARY STUDENTS							
Grade VIII	97	93–97	104–74	138–255	120–50	90–96	60
Grade VII	96	93–97	104–74	138–225	120–50	90–96	50–53

C. H. Stoelting Company. Complete outfit, including 25 copies of all printed material needed in quantity, and one copy of everything else, $77.25. The various items composing the series are also priced separately.

PUBLIC SPEAKING TEST
J. S. Gray (1926)

This measuring instrument perhaps scarcely deserves mention, yet because of the paucity of tests and scales for speech it is included here. It consists of a summary blank and an auditor's check blank. The first contains ten questions dealing with a speech and provides for rating according to five degrees of merit under each. These degrees are arranged in order and are each described in a few words. The auditor's check blank contains the same questions but has the descriptions of the degrees of merit arranged in random order. It is intended that a number of persons rate each speaker and that the results be summarized on the first blank described. Provision is made for combining the ratings into a single score.

The reliability coefficient of ratings is given as .78 and that with judges' ratings as .91. Norms are not given, but a table for changing ratings into per cent marks accompanies the test.

College Book Company. 35¢ per 50.

IV. Writing

It is even more true in the case of writing than in reading and spelling that the subject is not ordinarily taught as such in the

high school. It does, however, receive some attention in connection with English work, especially theme writing, and occasionally elsewhere, so that it seems well to mention briefly a few of the most satisfactory scales for measuring its merit. Perhaps the greatest value to be obtained from their use comes not from their employment by teachers for giving pupils' writing definite ratings, but rather from posting the scales on bulletin boards or otherwise making them readily available to pupils so that they can compare their own writing with them and attempt to reach whatever degree of merit has been set as the standard, if they are not already doing so.

The statement just made implies that standards of merit for pupils' writing have been set, and this is indeed the case. The usual method has been to derive standards from compilations of opinions as to the degree of proficiency pupils should reach. Those asked to express their opinions have in most cases been persons who employ clerks and others at work that requires a considerable amount of writing by hand and in which it is important that this writing be easily legible. Sometimes the average opinion of such a group has been raised somewhat before being set as the standard for writing when pupils graduate from the eighth grade or at any other fixed time. This is done to allow for probable deterioration.

In measuring writing most emphasis has been put on quality, although rate has not been neglected. Since rate of writing has no connection with a scale used in scoring quality, rate norms are not connected with particular scales. About eighty letters per minute appears to be the norm for pupils entering the secondary school. The statement of quality depends upon the scoring systems used in connection with the various scales.

If pupils are stimulated to write more rapidly than is their usual practice, quality becomes worse, whereas if they center attention on quality, speed is likely to be decreased. Therefore, if the directions given pupils when they are producing specimens to be rated overemphasize or underemphasize either element, the results cannot validly be compared on the basis of either quality or rate alone. For this reason attempts have been made to find a satisfactory method of combining rate and quality into a single score. Several suggestions have been made and urged, but none

has received general acceptance. For most purposes it is probably best to deal with the two phases of writing separately.

Studies of reliability indicate that there is comparatively little difference among the best scales in this respect. Moreover, they show that the correlation between ratings on different scales is frequently almost, if not quite, as high as that between two ratings according to the same scale. Coefficients of reliability ranging up to .95 or above have been reported, but about .75 may be taken as a fairly representative average figure. As with composition scales, so here also study of the scales and proper practice in their use increases reliability considerably.

Most of the scales provide directions for securing samples from pupils. Ordinarily these directions call for the writing of a portion of some well-known selection, such as "Mary Had a Little Lamb" in the lower grades and Lincoln's "Gettysburg Address" in the upper grades and high school, for a short period of time, perhaps two minutes. They usually provide for pupils to become familiar enough with the selection that they can almost, if not quite, write it from memory and, furthermore, that a copy be placed on the board so that little time will be lost in case pupils cannot recall it immediately. Although standardized directions must be followed if the scores given pupils' work are to be comparable with general norms, it is important from the standpoint of developing good writing habits that specimens be taken from pupils' ordinary writing when they do not know that they are to be rated. In other words, pupils should not be allowed to think that it is important to write well only on the special occasions when they know their performances are to be scored, but rather they should feel that any of their written work is likely to receive such ratings and, therefore, take reasonable care with all of it.

Scale for Handwriting of Children
E. L. Thorndike (1910)

This, the first standardized handwriting scale, shares the distinction with that of Ayres' of being one of the two that are much more widely known and employed than any others. It consists of specimens at each value or degree of merit from four

to eighteen, inclusive, there being only one specimen at each of the values from four to seven, and also at seventeen and eighteen, and from two to four at each of the others. The two or more specimens at one value represent different varieties of handwriting. The basis of rating was that of general merit. In general results or norms for high-school pupils have not been established, but apparently the average pupil at the time of entering high school writes about as well as value eleven or perhaps twelve, whereas in the opinion of competent persons he should write about thirteen or perhaps fourteen.

Bureau of Publications. 12¢ per copy, $10.00 per 100.

Reference: Thorndike, E. L. "Handwriting," *Teachers College Record*, 11:1–93, March, 1910.

Measuring Scales for Handwriting

L. P. Ayres (1912)

Probably the most widely used handwriting scale is that of Ayres, commonly known as the "Gettysburg Edition" because the specimens included deal with a selection from Lincoln's Gettysburg Address. It consists of a specimen of a dozen lines in length at each value by tens from twenty to ninety, inclusive. This is a revision of Ayres' earlier scale published in 1912, which differed chiefly in having three specimens, one in each of three styles of writing, at each value on the scale. It may still be desirable to use the earlier scale for rating the work of pupils the slant of whose writing is not similar to that in the later scale. The basis of determining the merit of the specimens was that of legibility, as determined by the rate at which the samples could be read by the judges. The data available appear to indicate that the Ayres scale results in slightly more reliable ratings than does that of Thorndike. The average quality of the writing of pupils entering high school is about sixty, whereas it is more or less generally agreed that seventy should be the standard.

Two years after constructing his original or three slant scale Ayres devised another for measuring adult handwriting, similar to it in general form and organization. Some persons have suggested that the use of this scale in high school is preferable since the writing of high-school pupils approaches that of adults

more closely than that of elementary children. To the writer this position scarcely appears valid, and apparently his view is sustained by persons who have employed one of Ayres' scales in the secondary school.

Publishing Department, Russell Sage Foundation. 10¢ per copy, $9.00 per 100, $75.00 per 1,000.

References: Ayres, L. P. "A Scale for Measuring the Handwriting of School Children," *Russell Sage Foundation Bulletin*, No. E 113. New York: Russell Sage Foundation, 1912. 16 p.

———. "A Scale for Measuring the Handwriting of Adults," *Russell Sage Foundation Bulletin*, No. E 138. New York: Russell Sage Foundation, 1915. 12 p.

Chart for Diagnosing Faults in Handwriting
F. N. Freeman (1914)

For diagnosing the most common faults that occur in handwriting this chart of Freeman's is probably the best instrument available. It is divided into five sections devoted, respectively, to uniformity of slant, uniformity of alinement, quality of line, letter formation, and spacing. Under each of these headings are two or more specimens of writing at qualities one, three, and five, with various helps, such as lines, arrows, and greatly enlarged writing, in determining the faults in the writing. The specimens have also been published in the form of five separate charts rather than combined on a single large sheet.

Houghton Mifflin Company. 30¢ per copy.

References: Freeman, F. N. "An Analytical Scale for the Judging of Handwriting," *Elementary School Journal*, 15:432, April, 1915.

———. *The Teaching of Handwriting*. Boston: Houghton Mifflin Company, 1915. 156 p.

Scale for Measuring Handwriting
Daniel Starch and C. T. Wise (1919)

This is a revision of an earlier scale, with the values of certain specimens more accurately determined and others which appeared unsatisfactory eliminated. It consists of samples of writing at each value from eleven to twenty-five, inclusive, and also

at values zero, two, four, five, seven, and nine. At all the values except four, five, twenty-four, and twenty-five, there are both representations of the capital letters, and a selection several lines in length. Although this scale contains specimens whose merit has been the most carefully determined of any of the existing scales, that of the others mentioned in this section has been well enough determined that its superiority in this respect is not great. One undesirable feature is that it is very large, being almost four feet in length and, therefore, cumbersome to use on a small desk or table. The average writing of pupils entering high school rates about seventeen or eighteen according to this scale, whereas the desirable minimum is probably about twenty. One defect of this scale is that it contains so many steps that the difference in merit between successive ones is in many cases too small to be readily observed.

Daniel Starch. 50¢ per copy, $2.75 per 6, $5.00 per 12.

References: Starch, Daniel. "A Scale for Measuring Handwriting," *School and Society*, 9:154–58, 184–88; February 1, 8, 1919.

———. "A Revision of the Starch Writing Scale," *School and Society*, 10:498–99, October 25, 1919.

BIBLIOGRAPHY

I. Reading and Vocabulary

Brinkley, S. G. "Relative Value of Different Types of Questions in Reading Tests," *School Science and Mathematics*, 25:703–8, October, 1925.

Burgess, M. A. *The Measurement of Silent Reading*. New York: Russell Sage Foundation, 1921. 163 p.

Foran, T. G. "The Present Status of Silent Reading Tests, Parts 1, 2," *Catholic University, Educational Research Bulletin*, Vol. 2, Nos. 2, 3. Washington, D. C.: Catholic Education Press, 1927. 27 and 51 p.

Gates, A. I. "An Experimental and Statistical Study of Reading and Reading Tests," *Journal of Educational Psychology*, 12:303–14, 378–91, 445–64; September, October, November, 1921.

———. *The Improvement of Reading*. New York: The Macmillan Company, 1927. 440 p.

———. "The Psychology of Reading and Spelling," *Teachers College, Columbia University, Contributions to Education*, No. 129. New York: Bureau of Publications, Teachers College, Columbia University, 1922. 108 p.

Gray, C. T. *Deficiencies in Reading Ability: Their Diagnosis and Remedies*. New York: D. C. Heath and Company, 1922. 420 p.

Gray, W. S. "Summary of Investigations Relating to Reading," *Supplementary Educational Monographs*, No. 28. Chicago: University of Chicago, 1925. 275 p.

———. "Summary of Reading Investigations (July 1, 1924, to June 30, 1928)," *Elementary School Journal*, 26:449–59, 507–18, 574–84, 662–73, February, March, April, May, 1926; 27:456–66, 495–510, February, March, 1927; 28:443–59, 496–510, 587–602, February, March, April, 1928; 29:443–57, 496–509, February, March, 1929.

Miles, D. H. "Significance of Reading in High School," *Contributions to Education*, Vol. I. Yonkers, New York: World Book Company, 1924, Chapter XXVIII.

Neher, H. L. "Measuring the Vocabulary of High School Pupils," *School and Society*, 8:355–59, September 21, 1918.

Newcomb, E. I. "A Comparison of the Latin and Non-Latin Groups in High School," *Teachers College Record*, 23:412–22, November, 1922.

Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High Schoo Instruction*. Yonkers, New York: World Book Company, 1927, p. 112–23

Smith, H. L. and Wright, W. W. "Reading," *Tests and Measurements*. New York: Silver, Burdett and Company, 1928, Chapter IX.

Starch, Daniel. *Educational Measurements*. New York: The Macmillan Company 1916 p. 38–40.

Stone, C. R. *Silent and Oral Reading*. Boston: Houghton Mifflin Company 1922. 306 p.

Symonds, P. M. *Measurement in Secondary Education*. New York: The Macmillan Company, 1927, p. 106–9.

Thorndike, E. L. *The Teacher's Word Book*. New York: Columbia University, 1921. 134 p.

———. "The Vocabularies of School Pupils," *Contributions to Education*, Vol. I. Yonkers, New York: World Book Company, 1924, Chapter VII.

II. Spelling

Andersen, W. N. "Determination of a Spelling Vocabulary Based upon Written Correspondence," *University of Iowa Studies in Education*, Vol. 2, No. 1. Iowa City: University of Iowa, 1921. 66 p.

Ashbaugh, E. J. "Iowa Spelling Scale for Grades II, III, IV, V, VI, VII, and VIII," *University of Iowa Extension Bulletins* No. 53, 54, and 55. Iowa City: University of Iowa, 1919. 20, 20, and 18 p.

Foran, T. G. "The Measurement of Ability in Spelling," *Catholic University, Educational Research Bulletin*, Vol. 1, No. 2. Washington, D. C.: Catholic Education Press, 1925. 37 p.

Gates, A. I. "The Psychology of Reading and Spelling," *Teachers College, Columbia University, Contributions to Education*, No. 129. New York: Bureau of Publications, Teachers College, Columbia University, 1922. 108 p.

Hollingworth, L. S. "Psychology of Special Disability in Spelling," *Teachers College, Columbia University, Contributions to Education*, No. 88.

New York: Bureau of Publications, Teachers College, Columbia University, 1918. 105 p.
Morton, R. L. "The Reliability of Measurements in Spelling," *Journal of Educational Method*, 3:321–28, April, 1924.
Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction*. Yonkers, New York: World Book Company, 1927, p. 108–12.
Smith, H. L. and Wright, W. W. *Tests and Measurements*. New York: Silver, Burdett and Company, 1928, p. 170–83.
Symonds, P. M. *Measurement in Secondary Education*. New York: The Macmillan Company, 1927, p. 95–96.
Tidyman, W. F. *The Teaching of Spelling*. Yonkers, New York: World Book Company, 1919. 178 p.
Witty, P. A. "Diagnosis and Remedial Treatment of Poor Spellers," *Journal of Educational Research*, 13:39–44, January, 1926.

IV. Writing

Freeman, F. N. *The Teaching of Handwriting*. New York: Houghton Mifflin Company, 1914. 156 p.
Gray, C. T. "A Score Card for the Measurement of Handwriting," *University of Texas Bulletin* No. 37. Austin: University of Texas 1915. 50 p.
Koos, L. V. "The Determination of Ultimate Standards of Quality in Handwriting for the Public Schools," *Elementary School Journal*, 18:423–46, February, 1918.
Pressey, L. C. and S. L. "Analyses of Three Thousand Illegibilities in the Handwriting of Children and of Adults," *Educational Research Bulletin (Ohio State University)*, 6:270–73, 285, September 28, 1927.
Pressey, S. L. and L. C. "Chart for Diagnosis of Illegibilities in Handwriting." Bloomington, Illinois: Public School Publishing Company, 1927.
Smith, H. L. and Wright, W. W. *Tests and Measurements*. New York: Silver, Burdett and Company, 1928, p. 184–92.
Thorndike, E. L. "Teachers' Estimates of the Quality of Specimens of Handwriting," *Teachers College Record*, 15:279–91, November, 1914.
West, P. V. "Improving Handwriting through Diagnosis and Remedial Treatment," *Journal of Educational Research*, 14:187–98, October, 1926.

CHAPTER VI

FOREIGN LANGUAGE

Introduction.—The place of foreign languages in the secondary school, the purposes to be attained by them, and the methods to be employed in teaching them have been the subject of a very great amount of controversy, especially in the last quarter of a century. This controversy has been concerned both with whether or not foreign language should be taught in high school at all, and also with the question of what language or languages should be taught. Because of this unsettled condition, which is certainly more pronounced than in the case of most secondary-school subjects, there is unusual need for the application of satisfactory methods of measurement. Before a beginning can be made in answering any of these questions it is important to determine what are the outcomes of instruction. It is useless to argue that foreign languages should be offered for certain reasons unless the purposes implied by these reasons are being attained, at least in some schools. Similarly such a question of method as whether the direct method should be employed because the chief outcome should be ability to speak the foreign language, or the indirect method because some other outcome is desired, cannot be answered unless the results of employing the two methods can be measured.

In general the problem of testing in the foreign languages is similar to that in English. Knowledge of vocabulary, ability to read, including oral and silent reading and translation, knowledge of grammar, and composition ability, may be named as the chief phases of the work that should be measured. So far not very much progress has been made in the construction of tests to measure oral reading and composition ability in foreign languages, but the other phases mentioned are fairly adequately dealt with in the several languages at all commonly offered in high school.

I. Latin

Although one of the outstanding results of the classical investigation was a list of objectives for the teaching of Latin,[1] there is not readily apparent on the surface a close connection between the abilities measured by most of the available tests and the statement of objectives just referred to. This is chiefly because the objectives tend to be ultimate instead of immediate, whereas the abilities tested are rather of the latter sort. In other words, such abilities as knowledge of vocabulary, of forms and syntax, ability to translate Latin into English and English into Latin, and so on, do not occupy a prominent place in the list of objectives, although the existing tests are almost entirely designed to measure such abilities.

Latin Vocabulary and Sentence Tests
V. A. C. Henmon (1917)
Tests 1, 2, 3, 4, X

These are among the earliest tests in this subject, and have been largely superseded, but still appear to merit inclusion here. The first four tests are of equal difficulty. Each consists of fifty Latin words arranged in scalar order, for which English words are to be given, and ten Latin sentences, likewise arranged in order of difficulty, to be translated. Test X contains only twenty-five words, arranged alphabetically, and twelve sentences. Before reaching the present form the tests passed through certain preliminary forms variously lettered and numbered. Twenty minutes are needed for each test.

Henmon selected thirteen recent or widely used beginners' books and determined all the words common to them, Cæsar, Cicero, and Virgil. From the 239 words thus chosen those used in the vocabulary tests were taken. The sentence tests were built up from the same words. By careful tryouts unsuitable sentences were discovered and either modified or eliminated, the difficulty of all the words and sentences carefully determined and the ma-

[1] "Aims or Objectives in the Teaching of Secondary Latin," *The Classical Investigation*, Part One. Princeton, New Jersey: Princeton University Press, 1924, Chapter III.

terials divided so as to form the four duplicate tests and also Test X.

The chief weakness of the tests is probably the fact that scoring is not highly objective. Since pupils are given freedom in translating the words and sentences, many responses occur concerning the correctness of which it is difficult to judge. This is especially true in the case of the sentences where it is very hard to decide whether or not slight errors are important enough to justify giving no credit, since scoring is on an all or none basis. A second objection is that because the weights of the items answered correctly are to be added, scoring is somewhat laborious. In addition to this, which may be called the approved way of securing scores, it is suggested that merely the number or the per cent right may be taken. It would seem that certainly for the vocabulary tests and perhaps for the others the number right would be the most satisfactory score. Reliability data are as follows:

Vocabulary:

$$r = .73,\ \text{P.E.}_{\text{meas.}} = 5.5,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .06,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .35.$$

Sentences:

$$r = .57,\ \text{P.E.}_{\text{meas.}} = 3.5,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .17,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .45.$$

A correlation of only .33 between vocabulary and sentences has been reported. Standard scores for the ends of the four years are given as follows:

	Tests 1–4				*Test X*			
	I	*II*	*III*	*IV*	*I*	*II*	*III*	*IV*
Vocabulary:								
Sum of scale values	71	84	95	97	—	—	—	—
Number right	33	39	44	45	13.5	18.5	22.0	23.0
Per cent right	66	78	88	90	55	74	88	90
Sentences:								
Sum of scale values	9.5	16.0	25.0	30.0	10	14	22	28
Number right	2.5	4.0	6.0	7.0	5.5	6.5	8.0	9.5
Per cent right	25	40	60	70	47	54	67	80

World Book Company. Specimen set 10¢; 50¢ per 25.

Reference: Brueckner, L. J. "The Status of Certain Basic Latin Skills," *Journal of Educational Research*, 9:390–402, May, 1924.
Henmon, V. A. C. "The Measurement of Ability in Latin," *Journal of Educational Psychology*, 8:515–38, 589–99, November, December, 1917; 11: 121–36, March, 1920.

Test in Latin Syntax

L. W. Pressey (1921)

Form 1

This test is composed of thirty-two English sentences each followed by four possible Latin translations of which only one is correct. The incorrectness of the others is due to the use of wrong forms and of incorrect vocabulary. Twenty minutes are allowed for the test and should be amply sufficient. The number of exercises in this test is too small to result in scores of much diagnostic value, but sufficient to yield a fair general measure of pupils' knowledge of some of the most common details of Latin syntax. Median scores for the ends of semesters are announced as follows:

Semester	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*
Median	16	14	19	20	23	26	26

Median and quartile scores based upon a total of almost six thousand cases, apparently well scattered over the country, have also been reported. These are given below:

Semester	*I*	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*
Third quartile	20	21	21	25	25	28	29	29
Median	15	16	16	20	20	23	26	25
First quartile	11	11	12	12	15	18	21	20

These indicate either that the test is not a satisfactory measure of pupils' ability or that their ability along the line of Latin syntax does not increase regularly from semester to semester. Probably the latter rather than the former is the true explanation. However, the differences from semester to semester are so small, due at least in part to the small number of items in the test,

that scores made thereon are scarcely satisfactory measures of progress.

Public School Publishing Company. Sample set 10¢; 50¢ per 25.

Reference: Brueckner, L. J. "The Status of Certain Basic Latin Skills," *Journal of Educational Research*, 9:390–402, May, 1924.

Test in Latin Verb Forms

Caroline Tyler and S. L. Pressey (1921)

Form 1

This is similar in form to the Pressey Latin Syntax Test, differing chiefly in that instead of English sentences with Latin translations it consists of Latin verb forms with four suggested English translations for each. Provision is made for a time limit of fifteen minutes. The medians given by the publishers are as follows:

Semester	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*
Median	17	20	19	20	24	27	26

Those from more than five thousand cases apparently well distributed differ somewhat and are as follows:

Semester	*I*	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*
Third quartile	20	22	24	24	27	28	31	30
Median	14	18	20	19	22	24	27	25
First quartile	9	13	16	15	17	19	22	21

Public School Publishing Company. Sample set 10¢; 50¢ per 25.

Reference: Brueckner, L. J. "The Status of Certain Basic Latin Skills," *Journal of Educational Research*, 9:390–402, May, 1924.

Latin Comprehension Test

B. L. Ullman and T. J. Kirby (1922)

Forms I, II

This test resembles a common type of reading test, as well as other foreign language tests. It contains ten Latin selections of

from three to six lines each, either selected from material commonly read by high-school pupils or similar to it, each followed by three or four English questions to be answered in English. The form was chosen after considering two other possibilities as being most acceptable for the purpose of testing ability to understand Latin. In the artificially constructed selections the vocabulary was limited to the words found by Henmon to be common to thirteen beginning books. As a result of trying out a much greater number of selections and questions on the selections, those finally included were chosen. Thirty minutes is the time. For single semester groups.

$$r = .62,\ \text{P.E.}_{\text{meas.}} = 2,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .10,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .41.$$

Apparently Form II is two or three points more difficult for first-year pupils than is Form I. Correlations averaging above .50 have been found between scores on this test and the New York Regents Examinations. Norms based on more than thirteen hundred pupils who have had one year of Latin, eleven hundred who have had two years, and between one and two hundred who have had three and four years are given as follows:

Year	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
IV	13	16	19	23	27	29	30
III	12	14	18	21	24	26	29
II	8	9	12	16	20	21	24
I	4	5	8	10	11	14	15

Bureau of Educational Research and Service. $1.75 per 100.

References: Byrne, Lee. "Latin Tests in Iowa High Schools," *University of Iowa Extension Bulletin*, No. 92. Iowa City: University of Iowa, 1923. 40 p.

Ullman, B. L. and Kirby, T. J. "A Latin Comprehension Test," *Journal of Educational Research*, 10:308–17, November, 1924.

Latin Composition Test

Edith R. Godsey (1922)

Forms A, B

Each of the three parts of the test consists of eleven English

sentences translated into Latin, but with four possible translations suggested for a certain word or larger portion of the sentence. Also there are fifteen or sixteen rules of syntax in each part. The numbers belonging to four of these rules appear in connection with each sentence. Pupils are to indicate the correct form and also the rule which applies to the construction. The syntax problems used were taken from Byrne's *The Syntax of High-School Latin*,[2] and all the vocabulary was based on Henmon's 239 word list. Provision is made for separate scores on the number of sentences or translations and of rules correct. The time is thirty minutes. The sentence scores have been found to correlate from .33 to .85 with vocabulary, translation, and syntax tests, whereas the rule scores showed very much lower correlations, in one or two cases even negative, with vocabulary and translation tests, but about as high, that is, from .50 to .75, with other tests on syntax and form. Norms based on about twenty thousand cases for the ends of the various semesters of high-school Latin are given as follows:

Semester	*I*	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*
Sentences	8	13	16	19	21	23	25	26
Rules	9	18	21	24	26	28	29	30

World Book Company. Specimen set 15¢; $1.00 per 25.

Reference: Brueckner, L. J. "The Status of Certain Basic Latin Skills," *Journal of Educational Research*, 9:390–402, May, 1924.

Latin Vocabulary Test

P. R. Stevenson (1923)

Forms 1, 2, 3

This test consists of three parts, each containing a column of forty English words, and another of twenty Latin words to be matched with twenty of the English words. Fifteen minutes are allowed, which seems short enough that speed should play a part in determining scores. However, the norms are high enough that apparently this is only slightly if at all the case. Three sets of

[2] Byrne, Lee, et al. *The Syntax of High-School Latin*. Chicago: University of Chicago Press. 54 p.

June norms are given, according to whether pupils have begun Latin in the seventh grade, the eighth, or the ninth. They are as follows:

Begun in grade	*Years studied*		
	I	*II*	*III*
IX	50	54	—
VIII	44	52	—
VII	43	53	58

Public School Publishing Company. Sample set 15¢; 50¢ per 25.

Latin Derivative Test

P. R. Stevenson and W. W. Coxe (1923)

Forms I, II, III

There are three matching exercises in this test, each composed of twenty-five Latin words and twenty English words derived from as many of them. In general the English words are not so difficult but that high-school pupils would know most of them, and a majority of the Latin ones are in most first-year books. The time is fifteen minutes. Tentative norms for the ends of semesters are:

Semester	*I*	*II*	*III*	*IV*
Norm	51	53	56	56

Evidently pupils make little progress along the line tested.

Public School Publishing Company. Sample set 15¢; 50¢ per 25.

Harvard Latin Tests

Alexander Inglis (1923)

Vocabulary, Morphology; Forms A, B, C, D, E of each;

Syntax;

Forms A, B

Each vocabulary test consists of 150 Latin words for which English meanings are to be given. Each syntax test is divided into two parts, of which the first is further subdivided. Part A

of the first contains twenty-three English sentences each containing an italicized noun or pronoun, for which pupils are to state the case, the construction, and the preposition, if any, used in Latin. Part B consists of twenty Latin prepositions, Part C, of twenty adjectives, and Part D, of sixteen verbs, after each of which the case or cases that it governs are to be named. In Part 2 there are twenty-nine English sentences with italicized verbs. Pupils are to give the correct mood, tense, construction, and conjunction, if any, which should be used for each verb in Latin. The morphology test contains about seventy items, some calling for single responses and others for several. Nominative singulars of nouns are given and a certain case and number of each called for, also various forms of nouns with the case, number and nominative singular of each required. For pronouns and adjectives a number of nominative singular masculines are given and particular cases, numbers, and genders called for, and also for some adjectives the derived adverbs with their comparatives and superlatives. For verbs certain forms and the principal parts are given, and the person, number, tense, mood, and voice of each demanded, whereas in other cases there are the principal parts and the specifications as to a form of each which pupils are to supply.

The vocabulary tests were constructed by sampling Lodge's *Vocabulary of High-School Latin*[3] in such a way as to include a certain number of words from each of six groups determined according to the frequency of occurrence. The more frequently words occurred, the greater the number of them included in the tests. Scale values were assigned according to the frequency of occurrence. The syntax tests are based upon Byrne's *Syntax of High-School Latin*,[4] and include in each form about half of the constructions employed in secondary-school Latin, exclusive of concordance, simple indicative, and word order. In this case also values were assigned according to frequency. The morphology test was constructed in similar fashion, being based upon the

[3] Lodge, Gonzalez. "The Vocabulary of High-School Latin," *Teachers College, Columbia University, Contributions to Education*, No. 9. New York: Bureau of Publications, Teachers College, Columbia University, 1907, 217 p.

[4] Byrne, Lee, et al. *The Syntax of High-School Latin*. Chicago: University of Chicago Press. 54 p.

frequency of occurrence of the various parts of speech with values assigned also in the same way as for the other tests. Each of the tests can easily be given in an ordinary high-school period of forty or forty-five minutes. In scoring one must add the scale values of the various items. So doing increases the work considerably over merely counting the number correct. The scale values are such that a perfect score on each test is one thousand points. The norms given below have been announced for pupils who have studied Latin various lengths of time. The vocabulary norms are based on a total of more than five thousand pupils from fifteen high schools, but the others are very tentative, being based upon only about three hundred cases.

	Months studied							*Years studied*							
	4	*4½*	*5*	*6*	*7*	*8*	*9*	*½*	*1*	*1½*	*2*	*2½*	*3*	*3½*	*4*
VOCABULARY															
Q_3	32	34	43	48	49	51	56	38	60	61	73	75	82	83	87
Md.	28	32	36	40	44	48	53	33	53	62	67	71	75	79	83
Q_1	24	26	30	34	40	42	45	26	50	56	62	68	69	75	78
SYNTAX															
Q_3	36	37	38	40	43	45	47	37	47	57	62	66	68	71	72
Md.	31	32	33	35	37	39	41	32	41	50	54	58	61	64	66
Q_1	2	4	6	10	15	19	23	4	23	42	48	52	54	55	57
MORPHOLOGY															
Q_3	24	25	26	28	30	32	34	25	34	42	55	69	76	78	81
Md.	11	12	13	15	17	19	21	12	21	30	42	54	63	70	77
Q_1	3	4	5	6	8	9	11	4	11	17	30	44	54	61	67

Ginn and Company. 80¢ per 30.

LATIN TEST
Dorrance S. White (1924)
Forms A, B

Part 1 is a vocabulary test containing a hundred Latin words with four suggested English meanings for each, and Part 2 a translation test of twenty sentences with either three or four English translations given for each. The words in Part 1 were selected according to the frequency with which they are found in the works of authors read for college entrance. Five were taken from those which occur one thousand times or more, twenty from the five hundred to one thousand group, fifty from the one hun-

dred to five hundred group, and twenty-five from the fifteen to one hundred group. They are arranged in this order, the more frequent coming first. The sentences were constructed so that the first ten should especially measure the ability of pupils in the first two years, the next five that of third-year pupils, and the last five that of those in the fourth year. Thirty-five minutes of working time are allowed. Data based on a small number of cases indicate low reliability, although examination of the test does not indicate that this would be expected.

$$r = .64,\ \text{P.E.}_{\text{meas.}} = 6,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .05,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .40.$$

Tentative norms based on about twelve hundred cases for the two parts are as follows:

Semester	*I*	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*
Vocabulary	29	39	48	56	63	70	76	82
Sentence	30	45	57	65	72	77	81	85

Percentile norms for 381 Pennsylvania seniors at the end of two years of Latin, 303 at the end of three and 3,177 at the end of four are:

Year	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
IV	92	101	120	140	159	173	180
III	70	77	98	117	138	154	163
II	61	68	83	99	117	132	144

World Book Company. Specimen set 20¢; $1.20 per 25.

Reference: Symonds, P. M. *Ability Standards for Standardized Achievement Tests in the High School.* New York: Bureau of Publications, Teachers College, Columbia University, 1927. 91 p.

Test in Latin Comprehension

R. J. Deferrari and T. G. Foran (1925)

This test contains nine Latin passages of from two to five lines each, either similar to or selected from the material com-

monly read in second, third, and fourth-year high-school Latin. Following each passage are from two to four questions in English, some of which call for double answers. The answers also are to be given in English. Pupils are allowed forty-five minutes and are instructed not to delay too long on the more difficult questions. Tentative norms of twenty-seven at the end of five semesters and thirty at the end of seven semesters of Latin have been announced. It appears that the scoring of this test is not entirely objective since in a number of cases it would easily be possible for pupils to give answers of doubtful correctness.

Catholic Education Press. $1.00 per 25¢; 15% discount to schools.

Test in Latin Vocabulary and Forms

R. J. Deferrari and T. J. Foran (1925)

Forms A, B

This test contains fifteen noun forms, ten pronoun, twenty verb, and five adjective forms, for each of which pupils are to give the meaning, and in addition for nouns the declension, number, gender, and case, for pronouns the kind, number and case, for verbs the conjugation, voice, mood, tense, person, and number, and for adjectives the gender, number, and case. The working time is thirty minutes. As is true in the case of the Latin Comprehension Test by the same authors, so in this test the vocabulary scores can scarcely be made perfectly objective. Pupils are allowed to write in their own answers, and some will occur concerning the correctness of which different scorers will disagree.

Catholic Education Press. $1.00 per 25; 15% discount to schools.

Progress Tests in Latin

B. L. Ullman and A. W. Smalley (1928)

This is a series of more than seventy tests dealing with vocabulary, sentences, forms, syntax, derivatives and word study, comprehension and Roman civilization. Each test calls for from ten to a hundred responses in single-answer, multiple-answer,

completion, and other forms. The tests are not standardized and probably never will be, but are intended rather to be used for practice and instructional purposes from day to day, or almost so. They are planned particularly for use with Ullman and Henry's *Elementary Latin*, but five recent textbooks in this field were considered in constructing them. A few of the items appear to be such as would not commonly be encountered in first-year work, but most of them are commonly met with there. The tests are combined into a booklet, but the leaves are perforated so that each test can, if desired, be removed easily. The time needed for each test is from fifteen to thirty minutes.

The Macmillan Company. 80¢ per set.

Latin Grammar Scales
M. E. Hutchinson (1928)
Forms A, B

These scales are each composed of thirty-five exercises in multiple-answer form. Part 1, which deals with nouns, pronouns, and adjectives, contains twenty-five exercises with weighted values running from 78 to 102, inclusive, by ones, and Part 2, on verbs, ten ranging from 89 to 100 by ones, but with none at 90 and 99. Each exercise is composed of an English sentence with four translations in Latin, only one of which is correct. The constructions included were chosen according to their presence in ten commonly used first-year books and their frequency in high-school Latin as given in Byrne's *Syntax of High-School Latin*.[5] The vocabulary employed is based upon Henmon's 239-word list. A coefficient of reliability of .85 has been found, but it appears to be for two preliminary forms of the test somewhat longer than the final ones. Except for the unnecessarily complicated method of determining scores, which is the same as that used in quite a number of tests, this scale appears to be a rather satisfactory measuring instrument. A score of a given amount indicates that a pupil who makes it can do tasks of that difficulty and get one-half of them correct. The time is twenty-five minutes.

[5] Byrne, Lee et al. *The Syntax of High-School Latin*. Chicago: University of Chicago Press. 54 p.

Public School Publishing Company. Sample set 15¢; 50¢ per 25.

Reference: Hutchinson, M. E. "A Standard Latin Grammar Test," *School and Society*, 27:47–48, January 14, 1928.

New York Latin Achievement Tests

H. G. Thompson and J. S. Orleans (1928)

Tests 1, 2; Forms A, B of each

These tests deserve to rank among the best in the subject of Latin. Test I is intended for first half-year students and Test 2 for second half-year students. The first consists of fourteen parts with from six to twenty-one items in each. They deal with vocabulary, both Latin-English and English-Latin, inflection, verb, noun, pronoun, and adjective forms, pronunciation, translation of single words not in their simplest form and of sentences, both Latin-English and English-Latin, and syntax. Several of these phases are tested in two or three different ways. Test 2 consists of the same number of parts, some of which are of exactly the same form as the corresponding parts in Test 1, and others of which differ somewhat. The time requirement for each of the tests is eighty minutes, but it is suggested that the first eight parts be given at one time and the last five at another. Thus either of the tests can conveniently be given in two ordinary high-school periods.

The tests are based upon the New York State Syllabus for first-year Latin. An examination of the tests indicates, however, that they are reasonably well suited to most first-year Latin courses, since they contain comparatively few items not found in the ordinary first-year textbook. Nowhere in the tests are there any suggested answers among which the pupils must choose, so that they resemble the ordinary examination more than do most standardized tests. Especially in vocabulary and translation, however, this results in responses that are difficult to score, and concerning the correctness of which different scorers will not always agree. Some of the other parts are such that there can be no other correct answers than those given in the key. Despite this element of subjectivity in the scoring, there are undoubtedly many teachers who prefer these tests to others

more highly objective because there are no suggested answers and pupils must rely entirely upon their own knowledge and initiative. Although Test 1 was tried out on about four thousand pupils, the published norms are based upon only a few hundred. For Test 2 they are based upon over sixty-six hundred. In both cases it appears that all the pupils were in the state of New York. The norms are as follows:

Percentile	*5*	*10*	*25*	*50*	*75*	*90*	*95*
Test 1							
First half-year	56	70	103	132	156	172	179
Second half-year	132	142	165	182	198	206	208
Test 2							
Second half-year	54	72	101	132	159	181	193

World Book Company. Specimen set 45¢, $1.30 per 25.

References: Orleans, J. S. and Thompson, H. G. "A Survey of Achievement in First Half Year Latin in New York State," *University of the State of New York Bulletin*, No. 892. Albany: University of the State of New York Press, 1927. 48 p.

———. "A Survey of Achievement in Second Half of First Year Latin in New York State," *University of the State of New York Bulletin*, No. 897. Albany: University of the State of New York Press, 1928. 24 p.

Latin Tests

Jessie D. Newby (1929)

First Year Tests 1, 2, 3, 4, 5, 6;

Second Year Tests 1, 2, 3, 4, 5, 6, 7, 8.

Each of the first-year tests and the first six of the second-year tests covers one six-weeks period of work. The last two of the second-year tests deal with Roman life and history and certain phases of mythology. The number of parts in the various tests varies from three to five and they contain from about sixty up to one hundred forty elements each, in exercises of several types. The time limit is forty minutes except for second-year Test No. 7, for which it is thirty, and No. 8, for which it is thirty-five.

Harlow Publishing Company. Sample set 10¢; 75¢ per 25, $2.50 per 100.

Diagnostic Tests in Latin Based on Gray and Jenkin's "Latin for Today: First Year Course"

F. N. Bacon (1922)

Tests I, II, III, IV, V, VI, VII, VIII, IX

Each of the nine tests of this series covers from eight to eleven lessons of the text mentioned above. There are from six to eight parts in each test dealing with as many different phases or topics. These differ somewhat, however, in the different tests. For example, the parts of Test II deal, respectively, with vocabulary, syntax, sentence comprehension, application of knowledge of Latin grammar to correct English, English derivatives, and history and life of the ancient Romans. The eight parts of Test IX cover vocabulary, indicative forms of all conjugations, fourth and fifth declensions, paragraph comprehension, English derivatives, facts of Roman legends, English and Latin suffixes, and application of Latin vocabulary to romance language vocabulary. Apparently pupils are allowed as much time as is necessary to complete the tests. In most cases this will probably be at least thirty minutes and often more.

This series appears to cover the content of Gray and Jenkin's book rather well and should be of decided value to teachers employing this text. With the exception of the portions dealing with Roman history, life and legend, the items included seem to fit most first-year books fairly well. However, since no norms have been announced nor any critical data published, the writer does not recommend these tests for general use by teachers using other texts.

Ginn and Company. $3.00 per 15 of all tests.

Latin Prognosis Test

J. S. Orleans and Michael Solomon (1926)

Form A

This test, which is intended for both high-school and college use, differs from those already described in its function. It does not measure knowledge of or ability in Latin, but capacity to learn Latin. There are nine subtests, each of which, from the

second to the seventh, inclusive, is preceded by a lesson covering the content dealt with by the subtest. The first subtest calls for English derivatives from Latin words of which the meanings are given. The last two deal with a vocabulary which the pupils had an opportunity to study in the very beginning of the test. The material dealt with by the lessons and tests includes singular and plural of nouns and verbs, gender, the nominative, dative, and accusative cases. Almost an hour is required to give the test, so that it is, unfortunately, too long for the ordinary high-school period. The average correlation found between scores on this test and the average of teachers' marks and Latin achievement-test scores is slightly above .80. In other words, in view of the known unreliability of school marks and the somewhat smaller but still present unreliability of standardized tests, this test appears to be about as valid for its purpose as could be expected. The correlation between scores on the test and Latin achievement is indicated by the following table in which the individuals concerned have been grouped in tenths on the two bases:

Standing in prognosis test	*Standing in achievement in Latin*									
	Tenths									
Tenths	*10th*	*9th*	*8th*	*7th*	*6th*	*5th*	*4th*	*3rd*	*2nd*	*1st*
1st					1	3	5	11	23	56
2nd			1	3	5	9	14	20	20	23
3rd		1	3	5	9	13	17	20	20	11
4th		3	5	9	14	16	18	17	14	5
5th	1	5	9	14	16	17	16	13	9	3
6th	3	9	13	16	17	16	14	9	5	1
7th	5	14	17	18	16	14	9	5	3	
8th	11	20	20	17	13	9	5	3	1	
9th	23	26	20	14	9	5	3	1		
10th	56	23	11	5	3	1				

World Book Company. Specimen set 15¢; $1.30 per 25.

II. Modern Foreign Languages

The three modern foreign languages most commonly taught in the secondary schools of this country, French, Spanish, and German, are not nearly so well provided with tests in so far as numbers are concerned as is Latin, but because of the high

merit of some of the tests, are fairly well taken care of in comparison with other subjects. Most of this development has come within a period of a very few years. Several French tests were made available by 1925 or thereabouts, but the best ones in this subject and practically all of those available in Spanish and German have appeared since that time. The reason for this is that the *Modern Foreign Language Study*[6] devoted considerable attention to the construction and standardization of tests in these three languages. This study commanded the services of experts both in test making and in language, provided clerical and other assistance, and rendered possible outstanding work in test construction. The American Council tests, which will be described in this section, were most closely connected with the study, but those of the Columbia Research Bureau and others were also employed.

The problems of testing in the modern languages are not different from those in Latin except that pronunciation and ability to translate into the foreign language hold more important places. Although not very much has been done along either of these lines, they have received somewhat more attention than in the case of Latin. On the whole, however, the varieties of exercises employed are very similar to those used in tests in the subject just mentioned, and also to many of those employed in connection with measurements in the English language.

French Tests

V. A. C. Henmon (1921)

Tests 1, 2, 3, 4

These are very similar to the same author's Latin tests. Each consists of sixty French words and twelve sentences, both arranged in order of difficulty. Pupils are to write correct translations. The words in twelve recent or widely used first-year texts were tabulated and from the resulting 448 common words those

[6] This study, which is probably the most comprehensive one ever made of school subjects in this country, is reported in a number of volumes. More than a dozen have appeared already and several others have been announced. Most of these will be found under their several authors in the bibliography at the end of this chapter. Those not included have been omitted because they touched very slightly, if at all, on measurement.

that appeared unsuitable were eliminated. The 240 words remaining were weighted according to the per cent of pupils giving the correct translations and then divided so as to form four sixty-word lists of equal difficulty. The sentences were formed from the 448 words, and after a similar trying out process were incorporated in the tests. Eight minutes for each vocabulary and twelve for each sentence test are the recommended time allowances, but it is stated that the tests are not speed tests.

From the standpoint of content these tests rank as high as any which deal with simple vocabulary and sentence translation. However, because pupils are given freedom in recording their responses, scoring cannot be as highly objective as is desirable. In many cases, especially as regards the sentences, it is doubtful whether pupils' responses should be scored as correct or incorrect, and teachers frequently disagree. Instructions are to score as right or wrong without partial credits, but there will be much difference of opinion as to how great the error in a sentence must be to make the sentence as a whole wrong. Provision is made for three methods of scoring, either the number of items correct, the sum of their scale values, or the per cent of items correct. Of these the second should perhaps be considered the approved method. Reliability is low.

$$r = .61,\ \text{P.E.}_{\text{meas.}} = 24,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .20,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .42.$$

Norms for the ends of the semesters are given as follows:

	Vocabulary					*Sentences*				
	I	*II*	*III*	*IV*	*VI*	*I*	*II*	*III*	*IV*	*VI*
Sum of weights	66	87	121	133	153	—	14	21	24	31
No. right	25	32	44	47	53	—	5.5	7.5	8.5	10
Per cent right	41	53	73	78	88	—	46	62	71	83

World Book Company. Specimen set 10¢; 50¢ per 25.

Reference: Henmon, V. A. C. "Standardized Vocabulary and Sentence Tests in French," *Journal of Educational Research*, 3: 81–105, February, 1921.

Harvard French Vocabulary Tests

Alice M. Twigg (1925)

Forms A, B

This test consists of a hundred French sentences in each of which one word is italicized, for which the English meaning is to be given, and fifty single French words not in sentences for which the meanings are likewise to be given. The basis for selecting the words included was a count of one hundred thousand word occurrences in standard French literature, scientific works, magazine articles, newspapers, and letters. A random selection of from fifteen to twenty-five words from each of seven frequency-of-occurrence groups is included in the test.[1] The words included were assigned weights or values according to their frequency so that the total of the transmuted weights is one thousand. It is stated that the tests should not require over thirty minutes, but this is not intended as a definite time limit, and pupils should be allowed to complete the test. Although a scoring key is provided, it is impossible to determine scores on an absolutely objective basis since in some cases responses will be given which are doubtful. The general instructions are to give credit if a pupil makes it clear that he knows the word, but this principle does not suffice to secure entire objectivity.

Coefficients of reliability ranging from .83 to .95 for single years and of .96 for all four years combined have been reported for this test. Correlations of from .65 to .80 for the single years were found with the vocabulary portion of the American Council Test, also correlations of from .36 to .67 with examination and semester marks, of from .36 to .41 with I. Q.'s, and of from .58 to .67 with a composite of I. Q.'s and semester marks. A comparison of this test with the American Council Vocabulary Test appears to indicate that its validity is superior. Median scores based on almost two thousand pupils in six schools are given as follows:

Semesters studied	*I*	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*
Median score	70	78	83	86	88	90	91	92

[1] In a few cases the principle of random selection was violated to avoid words that it was evidently undesirable to include.

Ginn and Company. 48¢ per 30.

Reference: Beatley, Bancroft. "A Comparative Study of Two Tests of French Vocabulary: The American Council French Test, and The Twigg French Vocabulary Test," in Henmon, V. A. C. *Achievement Tests in the Modern Foreign Languages. Publications of the American and Canadian Committees on Modern Languages*, Vol. 5. New York: The Macmillan Company, 1929, Appendix 2, p. 346–53.

AMERICAN COUNCIL FRENCH TESTS

ALPHA, Parts I, II; Forms A, B of each

V. A. C. Henmon, Algernon Coleman and M. R. Trabue (1925)

BETA, Forms A, B

Jacob Greenberg and B. D. Wood (1926)

GRAMMAR, Forms A, B

F. D. Cheydleur (1927)

This is by far the outstanding series of French tests and among the very best in any subject. It was very carefully developed as a part of the *Modern Foreign Language Study*. Part I of Alpha deals with vocabulary and grammar, Part II with silent reading and composition. Vocabulary is tested by a list of seventy-five French words for each of which the proper translation from among five English words is to be chosen. There are fourteen exercises on grammar calling for a total of fifty responses in various forms, such as completion, multiple-answer, and single-answer. For testing silent reading there are seven selections of from six to about twenty lines in length, each followed by four questions in French, to be answered in English. Composition is tested by presenting a picture and allowing eight minutes in which to write a composition about it in French.

Beta deals with vocabulary, comprehension, and grammar. The vocabulary portion is similar to that in Alpha, but contains one hundred instead of seventy-five words. Comprehension is tested by sixty incomplete French sentences for which the proper one of five expressions to complete the meaning is to be chosen. The grammar subtest calls for sixty responses. Short English sentences are given and in most cases all but one or two words of the correct French translations. In the case of a few very short sentences none of the French is given. In each case what

ever is lacking is to be supplied. The grammar test contains fifty multiple-answer exercises, of which one-third are concerned with verbs and one-fifth with pronouns. Each presents an English sentence with a portion of the French translation, and requires that the remainder be selected from among four possibilities. Each part of Alpha requires forty minutes of working time, Beta ninety minutes, and the grammar test twenty-two minutes.

Alpha is intended for high-school and college students, whereas Beta is for use in junior and senior high schools. The grammar test was prepared as a substitute for the grammar portion of Alpha, Beta, or other tests, for those who prefer the multiple-answer form. The words used in the Alpha vocabulary test were obtained by choosing one from each successive group of fifty words arranged in order of frequency as determined by Henmon's word count.[8] The Alpha grammar test deals with items common to current widely used texts. They were arranged in order of difficulty as determined by preliminary testing. The words in the Beta vocabulary test were selected mainly from among the most frequently used two thousand. On the whole these tests are the result of the most careful and comprehensive attempt to construct French tests ever put forth, the grammar test, for example, having been tried out on over twenty-five thousand high-school and college students. A number of studies of the reliability of the tests average about as follows:

	r	$P.E._{meas.}$	$\frac{P.E._{meas.}}{M}$	$\frac{P.E._{meas.}}{\sigma}$
ALPHA				
Vocabulary	.90	2	.06	.21
Grammar	.89	2	.09	.22
Silent reading	.83	1.5	.11	.28
Total score	.95	6	.09	.15
BETA				
Vocabulary	.94	1	.03	.16
Comprehension	.96	1	.04	.13
Grammar	.96	1	.05	.13
Total score	.97	3	.03	.12
GRAMMAR	.87	2	.06	.24

[8] Henmon, V. A. C. "A French Word Book Based on a Count of 400,000 Running words," *Bureau of Educational Research Bulletin*, No. 3. Madison: University of Wisconsin, 1924. 84 p.

The grammar test correlates about .83 with the grammar portion of Alpha and from .60 to .80 with the three other parts. The four parts of Alpha intercorrelate from .50 to .80 among themselves. Average correlations with school marks appear to be about .40, .55, .42 and .35, respectively.

For scoring compositions written as the fourth part of Alpha a French composition scale similar to English composition scales has been worked out by M. R. Trabue. One hundred twenty-five teachers of French gave their judgments in connection with its construction. The scale is presented in two parts, X containing samples at each even-numbered quality from two to sixteen, and Y at each odd-numbered one from three to seventeen. This unusual arrangement is employed to make the use of the scale easier to learn since differences between compositions of quality six and eight, for example, are much more easily observed than those between six and seven or seven and eight. After one becomes familiar with the scale it is recommended that it be employed as a unit.

The following norms for the various parts of Alpha are given. In most cases the high-school norms are based on enough individual scores to be fairly reliable. College norms are based on somewhat fewer cases, in a number of instances less than one

	Semester							
	I	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*
	High school							
Vocabulary								
Q_3	21	27	33	37	44	47	52	61
Md.	16	22	27	31	37	41	43	52
Q_1	11	17	22	28	31	36	37	47
Grammar								
Q_3	10	15	26	31	35	37	41	44
Md.	7	11	19	25	28	36	35	40
Q_1	5	7	12	17	21	27	27	33
Silent reading								
Q_3	8	11	16	17	20	21	21	24
Md.	6	8	13	15	17	19	19	21
Q_1	4	6	9	12	14	16	17	19
Composition								
Q_3	7.7	7.7	8.5	8.8	9.2	9.9	9.8	10.1
Md.	5.8	6.2	7.3	7.6	8.3	9.0	8.9	8.7
Q_1	4.1	5.0	5.8	5.9	6.9	7.8	8.5	7.4

						College		
Vocabulary								
Q_3	29	35	47	52	59	—	66	—
Md.	24	29	40	45	53	—	59	—
Q_1	19	24	34	35	46	—	55	—
Grammar								
Q_3	18	29	29	36	35	—	41	—
Md.	12	22	23	27	29	—	36	—
Q_1	9	14	17	21	22	—	30	—
Silent reading								
Q_3	16	17	21	21	24	—	24	—
Md.	13	15	19	21	21	—	23	—
Q_1	9	11	17	18	19	—	21	—
Composition								
Q_3	8.7	9.2	9.8	10.0	10.2	—	—	—
Md.	7.5	7.9	8.8	9.1	8.9	—	—	—
Q_1	6.1	6.3	7.4	7.9	7.6	—	—	—

hundred, but perhaps half of them are fairly well established.

The Beta results for each form are based upon almost nineteen thousand junior high-school pupils all in New York City. For regular classes they are as follows:

Semester	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
IX B	65	80	99	123	141	158	165
IX A	46	53	67	83	100	115	125
VIII B	33	38	47	58	70	83	91
VIII A	20	25	32	41	49	59	68

The norms for the Grammar Test are based upon from four hundred to almost three thousand cases in each semester. They are given on page 188. It will be seen that during the last three semesters high-school pupils tend to score higher than college students.

World Book Company. Alpha-Specimen set 35¢; $1.25 per 25 of either part. Beta-Specimen set 25¢; $1.30 per 25. Grammar-Specimen set 20¢; $1.25 per 25.

References: Breed, Frederick S. "The Reliability of the Trabue French Composition Scale," in Bagster-Collins, E. W. et al. *Studies in Modern Language Teaching. Publications of the American and Canadian Committees on Modern Languages*, Vol. 17. New York: The Macmillan Company, 1930, Chapter IV.

Cheydleur, F. D. "The American Council French Grammar Test, Selection

Semester	Percentile						
	5	10	25	50	75	90	95
COLLEGE							
VIII	47	51	60	69	80	89	94
VII	41	47	55	64	75	85	90
VI	40	43	51	62	70	77	83
V	41	45	53	61	70	77	82
IV	36	42	50	58	65	73	77
III	34	38	45	53	61	68	72
II	32	36	44	51	59	67	71
I	22	25	30	36	43	50	54
HIGH SCHOOL							
VIII	51	58	67	76	83	89	92
VII	39	49	60	71	79	86	90
VI	42	47	55	64	74	81	86
V	38	43	51	60	68	76	80
IV	30	34	43	52	62	71	76
III	25	29	36	45	55	65	70
II	18	21	27	36	45	55	62
I	11	15	20	27	35	43	49

Type. Preliminary Experiment at the University of Wisconsin," *Bureau of Educational Research Bulletin*, No. 8. Madison: University of Wisconsin, 1927. 35 p.

———. "The Construction and Validation of a French Grammar Test of the Selection or Multiple-Choice Type," *Journal of Educational Research*, 17:184–96, March, 1928.

Ford, H. E. "The Reliability of the Trabue French Composition Scale for Scoring Ten-Minute Compositions," in Bagster-Collins, E. W. *Studies in Modern Language Teaching. Publications of the American and Canadian Committees on Modern Languages*, Vol. 17. New York: The Macmillan Company, 1930, Chapter V.

Henmon, V. A. C. *Achievement Tests in the Modern Foreign Languages. Publications of the American and Canadian Committees on Modern Languages*, Vol. 5. New York: Macmillan Company, 1929. 363 p.

Wood, B. D. *New York Experiments with New-Type Modern Language Tests. Publications of the American and Canadian Committees on Modern Languages*, Vol. 1. New York: Macmillan Company, 1927, Parts 1 and 3.

Symonds, P. M. *Ability Standards for Standardized Achievement Tests in the High School.* New York: Bureau of Publications, Teachers College, Columbia University, 1927. 91 p.

Columbia Research Bureau French Test

A. A. Méras, Suzanne Roth, and B. D. Wood (1926)

Forms A, B (C, D also announced)

This undoubtedly deserves to rank close to the American

Council Tests in general merit. It is composed of three parts, dealing, respectively, with vocabulary, comprehension, and grammar. Part I contains one hundred French words for each of which the correct meaning is to be selected from among several English words, Part II requires that seventy-five French sentences dealing with content known to practically all high-school pupils be marked as true or false, and Part III includes one hundred completion exercises calling for one, two, or three words in each blank. The test is intended for use in both high school and college. The total working time is ninety minutes.

This test resulted from several years of work with three experimental editions that were tried out with large numbers of individuals. The vocabularies of twelve grammars and four composition books, which constituted a fair sampling of all those used in this country, were studied as a basis for Part I. The true-false statements in Part II and the completion statements in Part III are also largely based upon the results of the same study. Its reliability is high, as the following results show:

$$r = .96,\ \text{P.E.}_{\text{meas.}} = 4,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .025,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .13.$$

The intercorrelations between the three parts are above .70. Correlations with college marks average around .70, which is about as high as the unreliability of the marks allows.

June norms based on over thirteen thousand second-year, almost seven thousand third-year, and about five hundred fourth-year New York high-school pupils are given as follows:

Year	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
IV	159	168	188	207	224	238	244
III	138	149	168	185	201	215	224
II	98	109	130	152	172	190	201

World Book Company. Specimen set 20¢; $1.30 per 25.

References: Méras, A. A., Roth, Suzanne and Wood, B. D. "A Placement Test in French," *Contributions to Education*, Vol. 1. Yonkers, New York: World Book Company, 1924, Chapter XXV.

Wood, B. D. "The Regents Experiment with New-Type Examinations in

French, Spanish, German, and Physics, June, 1925," *New York Experiments with New-Type Modern Language Tests. Publications of the American and Canadian Committees on Modern Languages*, Vol. 1. New York: Macmillan Company, 1927, Part 2.

Silent Reading Test in French

M. E. Broom and L. P. Brown (1928)

Forms A, B

Each form consists of twenty short French paragraphs followed by two multiple-answer exercises likewise in French. Apparently the paragraphs and accompanying exercises are arranged at least roughly in order of increasing difficulty. A time limit of twelve minutes is provided which should result in the scores depending largely on speed as well as comprehension. The directions are such that the test is practically self-administering, merely requiring some one to time it. The coefficient of reliability is about .85. Tentative medians based on a total of almost four thousand individuals are reported as follows:

	High-school semester								*College year*		
	I	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*	*I*	*II*	*IV*
Score	9	16	19	22	26	30	33	36	18	31	37

Research Service Company. 2¢ per copy.

Standardized French Tests

É. B. de Sauzé (1929)

Grammar Tests, Numbers 1, 2, 3; Vocabulary and Comprehension Tests, Numbers 1, 2, 3

Each grammar test consists of fifty English sentences. Most of these are followed by the French translations with one or occasionally more omitted words to be supplied. A few are followed by several possible translations of which the correct one is to be indicated. The sentences employed and the words omitted are such that knowledge of various grammatical rules and usages rather than vocabulary is tested. Each of the vocabulary and comprehension tests consists of two parts. In the first part of

Test 1 are forty, and of Tests 2 and 3, sixty, French sentences, each of which contains an underlined word or phrase for which the meaning is to be given. In the second part are four short paragraphs in French, each followed by several English questions to be answered in English. The Number 1 tests cover the first nine lessons in *Cours Pratique en Français* by the same author, the Number 2 tests, the first twelve lessons, and the Number 3 tests, Lessons 12 to 26 in the same book.[9] Despite their adaptation to this book, the tests appear to contain items most of which are so common as to merit use elsewhere. Thirty-five minutes are allowed for each grammar test, thirty minutes for Vocabulary and Comprehension Test 1, forty-five for Number 2, and fifty for Number 3.

The following medians are based on about two thousand pupils, apparently all in Cleveland:

Tests	*1*	*2*	*3*
Vocabulary	19	18	16
Comprehension	27	26	28
Grammar	31	33	28

John C. Winston Company. 5¢ per copy.

Reference: de Sauzé, É. B. *The Cleveland Plan for the Teaching of Modern Languages with Special Reference to French.* Philadelphia: John C. Winston Company, 1929. p. 100–61.

Standard French Test

Peter Sammartino and C. A. Krause (1928)

Parts I, II, Form A

Each part contains a vocabulary, a grammar, and a comprehension subtest. The vocabulary subtests consist of fifty words each with five suggested meanings. In the next part are twenty-five short English sentences accompanied by their French translations with one blank in each. Comprehension is tested by two or three short French paragraphs followed by five English questions apiece. The vocabulary words were selected after a study of sev-

[9] Number 4 tests, covering Lessons 27 to 35, are in process of standardization.

eral compilations and the grammatical constructions after a study of a number of commonly-used texts. The total working time allowed is twenty-eight minutes for Part I and thirty-two for Part II. Reliability data are as follows:

	r	$P.E._{meas.}$	$\frac{P.E._{meas.}}{M}$	$\frac{P.E._{meas.}}{\sigma}$
Vocabulary	.86	4	.06	.25
Grammar	.81	5	.08	.30
Comprehension	.90	3	.04	.22

The following norms, based on a small number of pupils, are given for the preliminary test that included the material later divided into two parts.

	Percentile						
Semester	*5*	*10*	*25*	*50*	*75*	*90*	*95*
Vocabulary							
I	27	30	38	46	55	64	74
II	54	58	66	73	77	85	87
III	75	77	83	88	92	95	97
IV	71	78	85	91	94	98	99
V	88	90	92	94	97	99	99
Grammar							
I	7	11	16	27	37	46	53
II	26	28	37	46	54	62	66
III	48	56	63	70	78	84	88
IV	52	56	64	74	82	88	92
V	66	69	76	80	86	90	95
Comprehension							
I	28	34	42	53	64	76	80
II	62	67	74	81	85	92	96
III	68	73	82	88	94	98	99
IV	70	78	82	86	95	98	99
V	85	87	93	96	98	99	100

Public School Publishing Company. Sample set 15¢; $1.50 per 25.

References: Sammartino, Peter. "An Experiment in Modern Languages," *Bulletin of High Points*, 10:677–80, November, 1928.

———. "A Standardized Test in Modern Languages," *Journal of Educational Research*, 20:231–33, October, 1929.

Iowa Placement Examinations

FRENCH TRAINING

G. E. Vander Beke, G. D. Stoddard, C. E. Seashore, and G. M. Ruch (1924)

Forms A, B

A description of this test will be found along with that of the other Iowa Placement Examinations on page 380. It appears to deserve fairly high rank among French tests.

Audition-Pronunciation Tests in French

O. K. Lundeberg and J. B. Tharp (1929)

Forms A and B

Although this test is not yet available for general use, it seems to merit description here. The audition portion consists of three parts. The first, on phonetic accuracy, contains fifty sets of near homonymics. One of each is pronounced and pupils are to indicate which one of the four it is. Part II and III deal with aural comprehension. In Part II the tester reads a series of twenty incomplete statements in French and pupils write the missing words in English. In Part III twenty definitions in French are read for which the English words are to be supplied. The Pronunciation Test contains four parts. The first consists of twenty French words, each followed by four English words, of which the one containing the sound most similar to the French word is to be marked. In Part II are fifteen French expressions with certain letters in bold-face type. Pupils are to indicate the English letters which most nearly express the sounds of those in bold face. Part III requires the indication of the silent letters in twenty French words. Part IV has twenty word groups that are to be marked according to whether liaison should be made or not. Apparently the whole test requires about thirty to thirty-five minutes of actual working time. The average coefficients of reliability reported are about .85 and .80 for the Audition and Pronunciation Tests, respectively. Correlations of .60 to .70 with school marks, of .67 between audition and pronunciation, of .77 between audition and a routine class aural test, of .46 be-

tween audition and aural comprehension, and of .37 between the two measures of the latter have been obtained. Tentative norms are available, but will not be given here. Further experimentation upon the test and similar ones in Spanish and German is to be carried on.

O. K. Lundeberg, University of Illinois, or J. B. Tharp, Ohio State University

Reference: Tharp, J. B. "The Lundeberg-Tharp Audition-Pronunciation Test in French." To appear in an early issue of the *Modern Language Forum.*

Achievement Tests in Spanish

L. A. Wilkins (1925)

Tests I, II, III, IV; Forms A, B, C, of each

This series of tests has been prepared to accompany Wilkins' *New First Spanish Book.* Each covers approximately one-fourth of the book, or the work of a quarter year. There are nine subtests in each. These deal with pronunciation, syllabication, accentuation, vocabulary, idioms, grammar, verbs, comprehension, and composition. Several different types of exercises are employed. Because of their adaptation to a particular text and the fact that not only is there considerable difference in the content of first-year books, but also in the order of presenting what is included, it is probable that these tests should not be employed with other texts.

Henry Holt and Company. 10¢ per copy, $1.20 per 25.

American Council Spanish Tests

Alpha, Forms A, B

M. A. Buchanan, J. P. W. Crawford, and Hayward Keniston (1926)

Beta, Forms A, B

Frank Callcott, R. H. Williams, and B. D. Wood (1926)

Since these tests are similar in most points to the American Council French Tests, the description here will note only points of difference. The Alpha Vocabulary Test contains one hunderd

words selected on the basis of a word count of more than one million Spanish words.[10] In silent reading there are eight selections instead of seven. In the grammar portion of Beta there are sixty-five items instead of the sixty of the similar French subtest. The Composition Scale employed in connection with the latter part of Part II of Alpha was prepared by V. A. C. Henmon, and differs from the French Scale in that the twelve

ALPHA

	Semester							
	I	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*
				High school				
Vocabulary								
Q_3	31	35	44	51	59	59	66	74
Md.	23	28	37	44	50	49	60	66
Q_1	17	23	31	36	42	44	50	55
Grammar								
Q_3	14	17	23	25	30	30	38	40
Md.	10	13	16	18	24	24	31	33
Q_1	7	10	12	13	19	18	22	23
Silent reading								
Q_3	10	12	19	21	24	23	29	28
Md.	5	9	15	17	20	20	25	24
Q_1	3	6	11	13	16	16	21	20
Composition								
Q_3	4.8	4.8	4.9	5.3	5.8	6.0	6.9	7.4
Md.	3.6	3.7	3.8	4.5	4.6	4.8	5.5	5.9
Q_1	2.0	2.6	2.6	2.7	3.2	3.4	4.7	4.6
				College				
Vocabulary								
Q_3	41	51	59	63	72	80	—	—
Md.	34	41	51	54	66	70	—	—
Q_1	28	31	43	41	55	61	—	—
Grammar								
Q_3	20	28	30	29	35	41	—	—
Md.	15	18	23	23	27	37	—	—
Q_1	11	14	18	16	29	27	—	—
Silent reading								
Q_3	18	24	27	27	33	—	—	—
Md.	13	18	23	23	28	—	—	—
Q_1	9	13	18	19	25	—	—	—

[10] Buchanan, M. A. (Compiled by). *A Graded Spanish Word Book. Publications of the American and Canadian Committees on Modern Languages* Vol. 3. Toronto, Canada: University of Toronto Press, 1927. 195 p.

tween audition and aural comprehension, and of .37 between the two measures of the latter have been obtained. Tentative norms are available, but will not be given here. Further experimentation upon the test and similar ones in Spanish and German is to be carried on.

O. K. Lundeberg, University of Illinois, or J. B. Tharp, Ohio State University

Reference: Tharp, J. B. "The Lundeberg-Tharp Audition-Pronunciation Test in French." To appear in an early issue of the *Modern Language Forum*.

Achievement Tests in Spanish

L. A. Wilkins (1925)

Tests I, II, III, IV; Forms A, B, C, of each

This series of tests has been prepared to accompany Wilkins' *New First Spanish Book*. Each covers approximately one-fourth of the book, or the work of a quarter year. There are nine subtests in each. These deal with pronunciation, syllabication, accentuation, vocabulary, idioms, grammar, verbs, comprehension, and composition. Several different types of exercises are employed. Because of their adaptation to a particular text and the fact that not only is there considerable difference in the content of first-year books, but also in the order of presenting what is included, it is probable that these tests should not be employed with other texts.

Henry Holt and Company. 10¢ per copy, $1.20 per 25.

American Council Spanish Tests

Alpha, Forms A, B

M. A. Buchanan, J. P. W. Crawford, and Hayward Keniston (1926)

Beta, Forms A, B

Frank Callcott, R. H. Williams, and B. D. Wood (1926)

Since these tests are similar in most points to the American Council French Tests, the description here will note only points of difference. The Alpha Vocabulary Test contains one hunderd

words selected on the basis of a word count of more than one million Spanish words.[10] In silent reading there are eight selections instead of seven. In the grammar portion of Beta there are sixty-five items instead of the sixty of the similar French subtest. The Composition Scale employed in connection with the latter part of Part II of Alpha was prepared by V. A. C. Henmon, and differs from the French Scale in that the twelve

ALPHA

	Semester							
	I	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*
				High school				
Vocabulary								
Q_3	31	35	44	51	59	59	66	74
Md.	23	28	37	44	50	49	60	66
Q_1	17	23	31	36	42	44	50	55
Grammar								
Q_3	14	17	23	25	30	30	38	40
Md.	10	13	16	18	24	24	31	33
Q_1	7	10	12	13	19	18	22	23
Silent reading								
Q_3	10	12	19	21	24	23	29	28
Md.	5	9	15	17	20	20	25	24
Q_1	3	6	11	13	16	16	21	20
Composition								
Q_3	4.8	4.8	4.9	5.3	5.8	6.0	6.9	7.4
Md.	3.6	3.7	3.8	4.5	4.6	4.8	5.5	5.9
Q_1	2.0	2.6	2.6	2.7	3.2	3.4	4.7	4.6
				College				
Vocabulary								
Q_3	41	51	59	63	72	80	—	—
Md.	34	41	51	54	66	70	—	—
Q_1	28	31	43	41	55	61	—	—
Grammar								
Q_3	20	28	30	29	35	41	—	—
Md.	15	18	23	23	27	37	—	—
Q_1	11	14	18	16	29	27	—	—
Silent reading								
Q_3	18	24	27	27	33	—	—	—
Md.	13	18	23	23	28	—	—	—
Q_1	9	13	18	19	25	—	—	—

[10] Buchanan, M. A. (Compiled by). *A Graded Spanish Word Book. Publications of the American and Canadian Committees on Modern Languages* Vol. 3. Toronto, Canada: University of Toronto Press, 1927. 195 p.

BETA

Semester	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
IX B	55	65	83	103	124	150	178
IX A	40	46	58	72	87	101	121
VIII B	35	40	48	58	70	86	104
VIII A	21	25	33	43	55	73	93

specimens at values from zero to eleven, inclusive, are combined to form a single scale. The reliability of Alpha appears to be slightly lower than that of French Alpha; that of Beta approaches the corresponding French test very closely. The announced norms are based on from eight hundred down to less than a hundred pupils for the various semesters of high-school work, and from about six hundred to less than thirty in the case of college students.

World Book Company. Alpha-Specimen set 35¢; $1.25 per 25 of either part. Beta-Specimen set 25¢; $1.30 per 25.

References: Henmon, V. A. C. *Achievement Tests in the Modern Foreign Languages. Publications of the American and Canadian Committees on Modern Languages*, Vol. 5. New York: The Macmillan Company, 1929. 363 p.

Wood, B. D. *New York Experiments with New-Type Modern Language Tests. Publications of the American and Canadian Committees on Modern Languages*, Vol. 1. New York: The Macmillan Company, 1927, Parts 1 and 3.

Columbia Research Bureau Spanish Test

Frank Callcott and B. D. Wood (1926)

Forms A, B (C, D to appear later)

This test is similar to the Columbia Research Bureau French Test in form, method of construction, and many other characteristics. Its reliability is slightly higher, as follows:

$$r = .97,\ \text{P.E.}_{\text{meas.}} = 3,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .02,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .12.$$

Correlations of scores with old-type Regents Examinations average about .65 and intercorrelations between the parts about

.67. The reliability of Part II, however, is distinctly low, being only .68. Percentile norms for over two hundred second-year, two thousand third-year and five thousand fourth-year New York State pupils are:

Year	*Percentile*				
	10	*25*	*50*	*75*	*90*
IV	195	210	223	234	246
III	160	177	192	208	224
II	124	142	163	184	203

The mean scores for the three parts are also given, as follows:

Year	*II*	*III*	*IV*
Part I	70	81	91
Part II	52	58	64
Part III	43	54	70
Total	164	193	222

World Book Company. Specimen set 20¢; $1.30 per 25.

Reference: Wood, B. D. "The Regents Experiment with New-type Examinations in French, Spanish, German, and Physics, June 1925," *New York Experiments with New-Type Modern Language Tests. Publications of the American and Canadian Committees on Modern Languages*, Vol. 1. New York: The Macmillan Company, 1927, Part 2.

Test for Spanish Vocabulary

Maria S. Contreras, Eustace Broom, and Walter Kaulfers (1927)

Forms A, B, C

This test is rather unusual in its form. It consists of sixty-seven exercises in each of which are four Spanish and four English words, some one of which is similar in meaning to some one of the Spanish words. Evidently this form was adopted to lessen the chance of guessing, but it has been criticized as causing the test to be one of association rather than of vocabulary, and because pupils with extensive Spanish but limited English vocabularies make poor scores. This objection, how-

ever, may apply to the ordinary multiple-answer and certain other types of exercises as well as to the variety used in this test. The working time is twenty-two minutes. As a basis for selecting the words, those occurring in twenty or more books, at least five times in one of them, and in several vocabulary lists, were determined, and, after omitting a few derivatives and others which seemed unsuitable, tried out. Finally after testing six thousand pupils with each form three final lists of sixty-seven words each arranged in order of difficulty were constructed. Their average reliability is approximately as follows:

$$r = .90,\ \text{P.E.}_{\text{meas.}} = 2,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .03,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .20.$$

Tentative norms based on very few cases above the fifth semester in high school are given as follows:

Semester	*Percentile*				
	10	*25*	*50*	*75*	*90*
COLLEGE					
II	57	60	63	66	67
I	37	42	48	54	60
HIGH SCHOOL					
VII	59	62	64	65	66
VI	61	62	65	66	67
V	55	60	63	66	66
IV	48	55	59	65	66
III	46	51	57	61	65
II	32	39	45	53	58
I	18	27	37	48	57

Public School Publishing Company. Sample set 15¢; 75¢ per 25.

References: Broom, Eustace and Contreras, Maria S. "A Background Vocabulary List in Spanish," *Modern Language Journal*, 11:459–63, April, 1927.

Broom, Eustace, Contreras, Maria S., and Kaulfers, Walter. "A Test of Spanish Vocabulary," *High School Teacher*, 3:216–17, 234, June 1927.

Silent Reading Test in Spanish

Maria S. Contreras, Eustace Broom, and Walter Kaulfers

(1927)

Forms A, B

In this test are twenty Spanish paragraphs of from two to five lines each, each followed by multiple-answer exercises in Spanish. The paragraphs and exercises are based upon the same words as the vocabulary test by the same authors. The final form represents the results of trying out preliminary material on about five thousand pupils and apparently has been very carefully constructed. It has been criticized, however, on the ground that the time limit, which as twenty-two minutes, is so long that speed of comprehension is not measured, and also that in the case of a number of paragraphs knowledge of a comparatively few words therein may indicate the answer, whereas in the case of others general reading comprehension may not be measured, but rather knowledge of single words. Several also allow different interpretations which justify different answers. The average reliability reported from several sources is as follows:

$$r = .80,\ \text{P.E.}_{\text{meas.}} = 2,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .12,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .30.$$

Tentative norms are given below:

Semester	*Percentile*				
	10	*25*	*50*	*75*	*90*
College					
II	8	12	16	17	19
I	7	10	13	16	18
High School					
VIII	—	—	20	—	—
VII	15	16	17	17	19
VI	14	16	17	19	19
V	14	16	17	19	19
IV	13	14	16	18	19
III	11	13	15	17	19
II	8	10	13	14	16
I	5	7	9	12	14

Public School Publishing Company. Sample set 10¢; 50¢ per 25.

Reference: Broom, M. E. "A Silent Reading Test in Spanish," *Journal of Educational Research*, 16:357–64, December, 1927.

Stanford Spanish Tests

A. M. Espinosa and T. L. Kelley (1927)

Parts I, II, III; Forms A, B of each

Part I deals with grammar, Part II with vocabulary, and Part III with paragraph meaning. Two other parts dealing with sentence meaning and pronunciation are planned. The grammar test calls for fifty responses which indicate knowledge of verb and noun forms, pronouns, adjectives, and so forth. The vocabulary test is composed of eleven matching exercises calling for knowledge of a total of sixty-eight Spanish words. Part III contains ten short paragraphs in Spanish each of which is followed by five or six questions, also in Spanish, to be answered in English. The time limits are, respectively, twenty, fifteen, and twenty minutes, but apparently it is expected that many pupils will complete the parts in less time as provision is made for recording the number of minutes actually consumed. The tests are intended for both high-school and university use. For the three parts the measures of reliability for single year groups average about as follows:

$$r = .76,\ \text{P.E.}_{\text{meas.}} = 2.5,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .08,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .33.$$

For the whole test with all years combined the figures are:

$$r = .94,\ \text{P.E.}_{\text{meas.}} = 5,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .05,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .17.$$

Mean scores based on groups of from one to four hundred are given on the following page:

Norms are also available for three groups, A, B, and C, in each year's work, these groups being divided according to the amount of foreign language carried.

Year	*Grammar*	*Vocabulary*	*Paragraph meaning*
University			
II	35.5	50.0	38.0
I	31.7	40.0	29.4
High School			
IV	39.1	51.3	38.0
III	35.6	47.4	34.4
II	35.4	40.7	28.6
I	22.2	28.0	18.0

Stanford University Press. Sample set, 25¢; 80¢ per 25 of each part; $2.25 per 25 of all three parts.

Iowa Placement Examinations

SPANISH TRAINING

G. E. Vander Beke, G. D. Stoddard, C. E. Seashore, and G. M. Ruch (1924)

Forms A, B

A description of this test will be found along with that of the other Iowa Placement Examinations on page 380. It appears to deserve fairly high rank among Spanish tests.

American Council Alpha German Test

V. A. C. Henmon, B. Q. Morgan, Stella M. Hinz, C. M. Purin, and Elizabeth Rossberg (1925)

Forms A, B

This test is very similar to the American Council Alpha French and Spanish Tests. The vocabulary contains one hundred words selected from Kaeding's count based on almost eleven million running words.[11] The grammar test is similar to the French grammar test by Cheydleur, rather than to the one included in the American Council Alpha Test. The composition scale, which was constructed by Elizabeth Rossberg, resembles the Spanish more than the French one. All the specimens, fourteen, are combined into a single scale, sometimes known as

[11] Kaeding, F. W. *Häufigkeitswörterbuch der deutschen Sprache.* Berlin: 1898. 671 p.

Scale A. The first two are at qualities zero and .5, respectively, and the others at each unit quality from three up to fourteen, inclusive. The method of construction was the same as in French, the ratings of a total of seventy-six judges being used. Reliability data average about as follows:

	r	P. E. meas.	$\frac{P.E._{meas}}{M}$	$\frac{P.E._{meas.}}{\sigma}$
Vocabulary	.94	2	.06	.17
Grammar	.86	2	.10	.25
Silent reading	.87	2	.13	.24
Composition	.62	.7	.10	.41

Intercorrelations between the various parts range from .40 to .66, and correlations with school marks based, however, on a small number of cases, average about .40, .55, .55, and .40 for the four parts in order. Norms for six semesters of high-school and four semesters of college work are given. Those for high school are based on several hundred cases in each semester, and those for college on groups ranging from several hundred down to about one hundred.

	Semester					
	I	*II*	*III*	*IV*	*V*	*VI*
			High school			
Vocabulary						
Q_3	36	32	39	48	58	63
Md.	23	25	30	42	41	52
Q_1	11	15	18	18	29	40
Grammar						
Q_3	20	22	24	27	30	33
Md.	14	17	22	23	24	29
Q_1	11	13	16	16	15	22
Silent reading						
Q_3	14	16	20	23	24	29
Md.	8	12	13	17	19	24
Q_1	5	7	10	13	15	19
Composition						
Q_3	4.9	5.7	6.2	6.5	7.4	7.8
Md.	3.9	4.6	5.0	5.2	6.1	6.4
Q_1	3.2	3.8	4.1	4.2	4.8	5.3

	College					
Vocabulary						
Q_3	33	45	49	56	—	—
Md.	26	37	38	46	—	—
Q_1	20	27	29	36	—	—
Grammar						
Q_3	24	29	30	33	—	—
Md.	20	23	26	27	—	—
Q_1	17	19	21	23	—	—
Silent reading						
Q_3	22	25	29	31	—	—
Md.	16	19	24	26	—	—
Q_1	13	15	18	21	—	—
Composition						
Q_3	6.9	7.9	8.1	8.7	—	—
Md.	5.7	6.6	6.8	7.5	—	—
Q_1	4.8	5.6	5.7	6.0	—	—

World Book Company. Specimen set 40¢; Part I $1.30 per 25; Part II $1.25 per 25.

References: Henmon, V. A. C. *Achievement Tests in the Modern Foreign Languages. Publications of the American and Canadian Committees on Modern Languages*, Vol. 5. New York: The Macmillan Company, 1929. 363 p.

Rossberg, Elizabeth. "An Attempt to Establish a Standard Scale for German Composition." Doctoral dissertation. Madison: University of Wisconsin, 1926.

American Council on Education German Reading Scales

M. J. Van Wagenen and Sophia H. Patterson (1927)

Scale A, Divisions 1, 2

Division 1 is for the first and second years of German, and Division 2 for the second and third years. Each consists of fifteen paragraphs arranged in order of difficulty. Those in Division 1 range from value sixty-six to ninety-four, and those in Division 2 from seventy-six to one hundred four, by intervals of two in each case. Each paragraph is followed by from four to six statements dealing with its content which are to be marked as either true or false. Instructions are to collect all test booklets at the end of fifty minutes, although speed is not intended to be a factor in determining scores. The paragraphs

included were taken from standard works in German commonly read and after the tentative statements were formulated were read critically by a number of instructors in German. After trying them out with several hundred students, the most satisfactory paragraphs and statements were selected for inclusion in the scales. Provision is made for computing corrected scores and therefrom final scores by counting the number of errors made. The score given indicates that a student can read paragraphs and sets of statements of that difficulty and get one-half of them correct or its equivalent.

Public School Publishing Company. Sample set 20¢; 75¢ per 25.

Reference: Henmon, V. A. C. *Achievement Tests in the Modern Foreign Languages. Publications of the American and Canadian Committees on Modern Languages*, Vol. 5. New York: The Macmillan Company, 1929, p. 301–3.

Columbia Research Bureau German Test

C. M. Purin and B. D. Wood (1926)

Forms A, B (C, D announced)

These tests are similar in both method of derivation and numbers and types of exercises to the Columbia Research Bureau French Tests and, therefore, will not be described further. Correlations between scores on this test and marks based on old-time Regents Examination were found to be slightly higher than those in the case of the French Tests. The intercorrelations of the three parts ranged from .65 to .78, averaging the same as the French. Likewise the reliability was found to be almost identically the same. Norms for 1,482 second-year pupils, 667 third-year and 127 fourth-year, all in New York State, are given as follows:

Year	*Percentile*				
	10	*25*	*50*	*75*	*90*
IV	185	204	220	235	247
III	160	183	205	225	243
II	119	149	176	201	221

The mean scores for the three parts are as follows:

Year	*II*	*III*	*IV*
Part I	78	90	96
Part II	49	58	64
Part III	47	58	69
Total	172	203	228

Both sets of norms are based on about fifteen hundred second-year students, between six and seven hundred third-year students, and only a few over one hundred in the fourth year.

World Book Company. Specimen set 20¢; $1.30 per 25.

Reference: Wood, B. D. "The Regents Experiment with New-Type Examinations in French, Spanish, German, and Physics." *New York Experiments with New-Type Modern Language Tests. Publications of the American and Canadian Committees on Modern Languages*, Vol. 1. New York: The Macmillan Company, 1927, Part 2.

First Year German Objective Tests
Based on Vos's *Essentials of German*
O. J. Oswald (1928)

This is a series of about thirty tests topically arranged for use with first- and second-year German classes. Various types are used, such as matching, single-answer, completion, and so forth. Each test deals with one or, in a few cases, more than one, topic such as demonstrative and possessive pronouns, nouns, synopsis, proper nouns and time, relative interrogative pronouns, conditional sentences, passive voice, and so on. No answer key or table of norms is provided, and the tests may be regarded as for practice rather than measurement purposes.

Henry Holt and Company. 50¢ per set.

Prognostic tests.—Several attempts have been made to construct tests that would be prognostic of ability to learn modern foreign languages. In most cases the tests designed for this purpose are no longer available, apparently because they were not considered of sufficient merit to receive very wide use. Two

of the earlier attempts along this line seem to deserve brief mention, however. Handschin about 1918 prepared two so-called "Pre-Determination Tests" [12] which, however, apparently never passed beyond mimeographed form. They were more or less similar to the Orleans-Solomon Prognosis Test in Latin in their general plan, presenting material to be studied and then testing how well it had been mastered. They dealt with Esperanto, apparently on the supposition that if pupils showed ability in learning that language they would also be able to learn the modern languages commonly taught in high school.

Another test of the same general sort was the Wilkins "Prognosis Test in Modern Languages," [13] which was for several years commercially available, being published by the World Book Company. It was similar to Handschin's tests in presenting material to both eye and ear, but differed in employing French and Spanish rather than Esperanto. There were two parts of which the second was to be given after four weeks' study of the language.

Probably the best test of this sort that is at present available is the Foreign Language Aptitude Test of the Iowa Placement Examinations. This was constructed by G. D. Stoddard and G. E. Vander Beke and deals entirely with the ability to learn Esperanto. Since the whole Iowa Placement series is described in this book, nothing further concerning this test will be included at this point. Two other such tests are described below.

Foreign Language Prognosis Test
P. M. Symonds (1928)
Tests A, B

Test A consists of four parts dealing with English inflection, word translation from English into Esperanto, sentence translation from Esperanto into English, and word translation from Esperanto. The parts of Test B deal with the formation of English parts of speech, word translation from Esperanto into

[12] Handschin, C. H. "Tests and Measurements in Modern Language Work," *Modern Language Journal*, 4:217–25, October, 1920.

[13] Wilkins, L. A. "Results in a Prognosis Test Given to Pupils Beginning French and Spanish," *Bulletin of High Points*, Vol. I, No. 8. New York: Board of Education, 1919, p. 26–30.

English, artificial language, and sentence translation from Esperanto into English. Test A calls for 115 and B for 125 responses in various forms. In general the material dealing with Esperanto involves the giving of examples, especially vocabulary, to be studied and then the testing of how well they have been mastered. Each test requires forty-four minutes of working time. Although not duplicate forms, A and B are of approximately equal validity and difficulty.

These tests were constructed after a considerable amount of experimental work by Symonds in attempting to predict foreign-language success. About half of the material included was actually tried out, whereas the remainder was constructed in accordance with Symonds' experience in this work. It appears that if the different subtests are properly weighted the multiple correlation with success in French and Spanish is somewhat above .60. The weights, however, differ decidedly for the two subjects.

Bureau of Publications, Teachers College. Sample set 10¢.

Reference: Symonds, P. M. "A Foreign Language Prognosis Test," *Teachers College Record* 31:540–56, March, 1930.

———. "A Modern Foreign Language Prognosis Test," in Henmon, V. A. C., et al. *Prognosis Tests in the Modern Foreign Languages. Publications of the American and Canadian Committees on Modern Languages*, Vol. 14, New York: Macmillan Company, 1929, Chapter VI.

Modern Language Prognosis Test

M. A. Luria and J. S. Orleans (1930)

Form A

This test is so similar to the Orleans-Solomon Latin Prognosis Test that no detailed description will be given. It is intended, the authors state, to test ability to learn French, Spanish, or Italian, however only French and Spanish are employed. A correlation of .68 between scores and those on achievements tests was obtained with a preliminary shorter form. It is estimated that for Form A the corresponding correlation is about .75. No norms are given, but the following table similar to that for the Latin Test showing comparative standing on this test and in modern language achievement is used.

Standing in Prognosis Test	*Standing in Achievement in Modern Language*									
	Tenths									
	10th	*9th*	*8th*	*7th*	*6th*	*5th*	*4th*	*3rd*	*2nd*	*1st*
Tenths										
1st				1	2	4	7	12	23	51
2nd		1	2	4	6	9	14	18	24	23
3rd		2	4	7	10	13	16	18	18	12
4th	1	4	7	10	12	15	16	16	14	7
5th	2	6	10	12	14	15	15	13	9	4
6th	4	9	13	15	15	14	12	10	6	2
7th	7	14	16	16	15	12	10	7	4	1
8th	12	18	18	16	13	10	7	4	2	
9th	23	24	18	14	9	6	4	2	1	
10th	51	23	12	7	4	2	1			

World Book Company. Specimen set 15¢; $1.30 per 25.

BIBLIOGRAPHY

I. Latin

Allen, W. S. "A Study in Latin Prognosis," *Teachers College, Columbia University, Contributions to Education*, No. 135. New York: Bureau of Publications, Teachers College, Columbia University, 1923. 40 p.

Briggs, T. H. "Prognosis Tests of Ability to Learn Foreign Languages," *Journal of Educational Research*, 6:386–92, December, 1922.

Brown, H. A. "A Study of Ability in Latin in Secondary Schools." Oshkosh, Wisconsin: State Normal School, 1919. 170 p.

Breuckner, L. J. "The Status of Certain Basic Latin Skills," *Journal of Educational Research*, 9:390–402, May, 1924.

Byrne, Lee. "Latin Tests in Iowa High Schools," *University of Iowa Extension Bulletin*, No. 92. Iowa City: University of Iowa, 1923, 40 p.

Clem, O. M. "Detailed Factors in Latin Prognosis," *Teachers College Columbia University, Contributions to Education*, No. 144. New York: Bureau of Publications, Teachers College, Columbia University, 1924. 52 p.

———. "Latin Prognosis: A Study of the Detailed Factors of Individual Pupils," *Journal of Educational Psychology*, 16:160–69, March, 1925.

Hanus, P. H. "Measuring Progress in Learning Latin," *School Review*, 24: 342–51, May, 1916.

Jordan, J. N. "Prognosis in Foreign Language in Secondary Schools," *School Review*, 33:541–46, September, 1925.

Lohr, L. L. "A Latin Form Test for Use in High School Classes," *High School Journal*, 1:7–9, 14–17; November, December, 1918.

———. "A Latin Form Test," *High School Journal*, 5:217–23, December, 1922.

Starch, Daniel. *Educational Measurements.* New York: The Macmillan Company, 1916, p. 171–76.

————. "A Test in Latin," *Journal of Educational Psychology,* 10: 489–500, December, 1919.

Wentworth, M. M. "An Experiment with Two Latin Tests," *Division of Educational Research, School Document* No. 26. Los Angeles: City School District, 1919.

Woody, Clifford. "Report of the Latin Investigation in Various High Schools of Michigan," *Bureau of Educational Reference and Research Bulletin,* No. 64. Ann Arbor: University of Michigan, 1924. 40 p.

————. "The Ullman-Kirby and Godsey Latin Tests and the Carr English Vocabulary Test," *Bureau of Educational Reference and Research Bulletin,* No. 56. Ann Arbor: University of Michigan, 1923.

The Classical Investigation, Part 1. Princeton, New Jersey: University Press, 1924. 305 p.

Modern Foreign Languages

Bagster-Collins, E. W. et al. *Studies in Modern Language Teaching. Publications of the American and Canadian Committees on Modern Languages,* Vol. 17. New York: The Macmillan Company, 1930. 491 p.

Boveé, A. G. "A Suggested Score Card for Attainment in Pronunciation," *Modern Language Journal,* 10:15–19, October, 1925.

Briggs, T. H. "Prognosis Tests of Ability to Learn Foreign Languages," *Journal of Educational Research,* 6:386–92, December, 1922.

Buchanan, M. A. (Compiled by). *A Graded Spanish Word Book. Publications of the American and Canadian Committees on Modern Languages,* Vol. 3. Toronto: University of Toronto Press, 1927. 195 p.

Buchanan, M. A., and MacPhee, E. D. *An Annotated Bibliography of Modern Language Methodology. Publications of the American and Canadian Committees on Modern Languages,* Vol. 8. Toronto: University of Toronto Press. 428 p.

Buswell, G. T. *A Laboratory Study of the Reading of Modern Foreign Languages. Publications of the American and Canadian Committees on Modern Languages,* Vol. 2. New York: The Macmillan Company, 1927. 100 p.

Cheydleur, F. D. (Compiled and Edited by). *French Idiom List. Publications of the American and Canadian Committees on Modern Languages,* Vol. 16. New York: The Macmillan Company, 1929. 154 p.

Coleman, Algernon. *The Teaching of Modern Foreign Languages in the United States. Publications of the American and Canadian Committees on Modern Languages,* Vol. 12. New York: The Macmillan Company, 1929. 299 p.

Deihl, J. D. "The Basis of Educational Tests in Modern Foreign Languages," *Modern Language Journal,* 7:269–73, March, 1923.

Handschin, C. H. "Tests and Measurements in Modern Language Work," *Modern Language Journal,* 4:217–25, February, 1920.

Hauch, E. F. (Compiled by). *German Idiom List. Publications of the Amer-*

ican and Canadian Committees on Modern Languages, Vol. 10. New York: The Macmillan Company, 1929. 98 p.

Henmon, V. A. C. *Achievement Tests in the Modern Foreign Languages. Publications of the American and Canadian Committees on Modern Languages*, Vol. 5. New York: The Macmillan Company, 1929. 363 p.

————. "A French Word Book Based on a Count of 400,000 Running Words," *Bureau of Educational Research Bulletin*, No. 3. Madison: University of Wisconsin, 1924. 84 p.

————. *Prognosis Tests in the Modern Foreign Languages. Publications of the American and Canadian Committees on Modern Languages*, Vol. 14. New York: The Macmillan Company, 1929. 182 p.

Jordan, J. N. "Prognosis in Foreign Language in Secondary Schools," *School Review*, 33:541–46, September, 1925.

Kaeding, F. W. "*Häufigkeitswörterbuch der deutschen Sprache*, Berlin: 1898. 671 p.

Keniston, Hayward (Compiled by). *Spanish Idiom List. Publications of the American and Canadian Committees on Modern Languages*, Vol. 11. New York: The Macmillan Company, 1929. 108 p.

Lundeberg, Olav K. "Recent Developments in Audition Speech Tests," *Modern Language Journal*, 14:193–202, December, 1929.

Lundeberg, O. K. "Testing Achievement in Foreign Language as to Speech and Audition," *High School Conference Proceedings*, 1928. Urbana: University of Illinois, 1929, p. 238–43.

Morgan, B. Q. *German Frequency Word Book. Publications of the American and Canadian Committees on Modern Languages*, Vol. 9. New York: The Macmillan Company, 1928. 87 p.

Russell, G. O. "A Silent Reading Test," *Modern Language Journal*, 9:459–68, May, 1925.

Starch, Daniel. *Educational Measurements*. New York: The Macmillan Company, 1916, p. 177–87.

Uhlendorf, B. A. "Prognosis Tests for Students of Foreign Languages," *High School Conference Proceedings*, 1922. Urbana: University of Illinois, 1923, p. 309–12.

Vander Beke, G. E. (Compiled by). *French Word Book. Publications of the American and Canadian Committees on Modern Languages*, Vol. 15. New York: The Macmillan Company, 1929. 188 p.

Ward, C. F. *Minimum French Vocabulary Test Book*. New York: The Macmillan Company, 1926. 101 p.

Wilkins, L. A. "Results in a Prognosis Test Given to Pupils Beginning French and Spanish," *Bulletin of High Points*, 1:26–30, October, 1919.

Wood, B. D. *New York Experiments with New-Type Modern Language Tests. Publications of the American and Canadian Committees on Modern Languages*, Vol. 1. New York: The Macmillan Company, 1927. 339 p.

CHAPTER VII

MATHEMATICS

Introduction.—This chapter will be devoted to tests in the usual mathematical subjects pursued in high school, that is, algebra and geometry, with also a little attention given to trigonometry and general mathematics. Arithmetic is sometimes offered in high school, but as it is distinctly an elementary-school subject, no discussion of it will be included here. Also commercial arithmetic tests will not be mentioned in this chapter, but rather in the chapter dealing with commercial tests.

Algebra and geometry may be ranked among the high-school subjects that are relatively well taken care of by standardized tests. Both were among the first of the high-school subjects in which such tests were constructed, and although there appear to have been a number of years in which there was little interest in geometry tests, yet by the date of writing both subjects have received a considerable amount of attention along this line. The Report of the National Committee on Mathematical Requirements[1] was undoubtedly a source of considerable stimulation, though not probably to the same extent that the *Classical Investigation* was in Latin and the *Modern Foreign Language Study* in French, German, and Spanish. Perhaps there would have been even more interest in standardized tests in high-school mathematics had it not been for the rather general belief that the ordinary teachers' tests in this subject were relatively objective since in mathematics almost everything is definitely right or wrong. Although this is in a certain sense true, especially in the case of algebra, yet studies of school marks referred to elsewhere in this volume[2] indicate that in actual practice the marks given pupils' mathematics papers by their

[1] Young, J. W. (Chairman). *The Reorganization of Mathematics in Secondary Education.* The Mathematical Association of America, 1923. 652 p.

[2] See page 6.

teachers are about as unreliable as those given in any other subject. Even though there is little or no doubt as to what the correct answer to an example or problem is, yet such factors as whether an answer is to be reduced to its lowest or simplest terms or not, whether partial credit is to be allowed if a portion of the work is correct or none at all unless it is entirely correct, whether allowance is to be made for what are evidently careless slips when the pupil knows better, and so on, result in decided differences in the marks teachers actually assign the same work. There is, therefore, just as great a need for objective tests and scoring methods in mathematics as in any other subject.

Another hindrance to the development of tests in this field is that there has been much uncertainty as to just what the content of high-school courses therein should be. How far the mechanics of computation should be emphasized and how far reasoning problems, whether what is sometimes called the mathematics of daily life or more theoretical mathematics largely preparatory for more advanced work in the subject should be offered, and other similar questions have not received nearly unanimous answers. It has, therefore, been somewhat difficult to select large numbers of items common to all or practically all textbooks and courses of study. It is, of course, true that more or less the same condition prevails in all subjects, but apparently it is somewhat worse in the field of mathematics than in many others. Despite these handicaps, however, a number of tests of real merit have been produced and will be described in the following pages.

I. Algebra

The beginning of standardized tests in algebra appears to have been a year or two earlier than in geometry and two rather well-known series of tests appeared at a comparatively early date. From then on the development has been fairly steady. The general tendency of most tests has been to include exercises representative of the important topics and procedures dealt with. In a very few cases, however, tests have been designed to measure pupils' ability along some one or a very few lines on the theory that these were the important outcomes of high-

school algebra and that, therefore, pupils' general achievements and ability in the subject could be best tested by concentrating on the important outcomes. Of these the ability to solve equations is probably outstanding.

First-Year Algebra Scales
H. G. Hotz (1918)
Series A, B

Each series contains five scales dealing with the following topics: addition and subtraction, multiplication and division, equation and formula,[8] problems, graphs. The first three scales of Series B each contains about twenty-five exercises, the last two only about half as many. Those in Series A contain from eight to twelve exercises, selected from among those in B. Exercises are arranged in order of increasing difficulty, thus measuring power rather than speed. Forty minutes is the time for each of Series B except the graph scale, and twenty minutes for the first three of Series A. The graph scale and the Series A problem scale are allowed twenty-five minutes.

These scales, which are among the earliest and best known for high-school use, were carefully derived. Preliminary forms were given, revised, and tried out again. In all about 16,000 pupils in eleven different states participated in the standardization. Although quite a number of other algebra tests have appeared, these are still receiving some use, but it is doubtful if they should now be ranked as among the best. They have been criticized as not conforming to the educational objectives of algebra, although they are intended to cover practically all that is commonly taught in first-year algebra, and Hotz states that he "firmly believes that no essential algebraic process has been ignored." Reliability data for Series A are:

$$r = .92,\ \text{P.E.}_{\text{meas.}} = 2,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .07,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .20.$$

For Series B the reliability is somewhat higher since the number of exercises is greater. Hotz's median standards are as follows:

[8] If only a single scale is to be used, this one is recommended.

Scale	*Series A*			*Series B*		
	Months studied			*Months studied*		
	3	*6*	*9*	*3*	*6*	*9*
Addition and subtraction	5.0	6.8	7.9	9.7	12.9	14.4
Multiplication and division	5.3	6.3	7.9	9.6	14.0	16.3
Equation and formula	4.9	7.1	7.8	7.8	14.3	16.0
Problem	4.3	4.9	5.6	5.4	6.5	7.5
Graph	2.8 *		5.6	3.7 *		7.2

* For four and one-half months.

In addition to these general standards, medians for schools of different sizes, different states and cities, and for schools using different textbooks are available.

Bureau of Publications. Specimen set 75¢; 70¢ per 100; except Graph Scale, $1.25 per 100; manual, 75¢ per copy.

References: Eells, W. C. "Hotz Algebra Scales in the Pacific Northwest," *Mathematics Teacher*, 18:418–27, November, 1925.

Harris, Eleanora and Breed, F. S. "Comparative Validity of the Hotz Scales and the Rugg-Clark Tests in Algebra," *Journal of Educational Research*, 6:393–411, December, 1922.

Hotz, H. G. "First Year Algebra Scales," *Teachers College, Columbia University, Contributions to Education*, No. 90. New York: Bureau of Publications, Teachers College, Columbia University, 1918. 87 p.

Symonds, P. M. *Ability Standards for Standardized Achievement Tests in the High School*. New York: Bureau of Publications, Teachers College, Columbia University, 1927. 91 p.

Algebra Test

L. L. Thurstone (1919)

This test, which is no longer receiving much use, is briefly described along with others of the series of Vocational Guidance Tests on page 377.

Illinois Standardized Algebra Tests

W. S. Monroe and L. W. Williams (1920)

This is a revision of, or substitute for, the earlier Monroe Standard Research Tests in Algebra. It included four subtests, each of which contains twenty exercises. All deal with the

simple equation. The first subtest has both unknowns in the first member of the equation, the other subtests have one in each. Subtest III involves parentheses and Subtest IV fractions. Within each subtest are five cycles of four exercises each, the difference consisting in the arrangement of signs. Pupils' actual working time is twenty-eight minutes. Evidently the test tends to measure speed to a considerable degree, since the median number of attempts even for third-semester pupils does not run over two-thirds of the number of exercises. As measures of the simple equation scores on this test seem to possess considerable validity. In so far as the simple equation is the fundamental topic of first-year algebra, scores on the test are valid as measures of algebraic ability in general. The cycle arrangement, with several examples of each subtype, gives the results considerable diagnostic value.

$$r = .88,\ \text{P.E.}_{\text{meas.}} = 2.2,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .10,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .23.$$

The following tentative norms for the ends of semesters give both attempts and rights:

Semester	*I*		*II*		*III*	
Test	*Att's.*	*R'ts.*	*Att's.*	*R'ts.*	*Att's.*	*R'ts.*
I	9.8	5.0	10.5	6.4	11.7	8.7
II	10.8	4.6	11.8	6.4	12.6	8.3
III	11.0	3.6	12.2	5.5	13.9	7.3
IV	8.8	1.0	11.3	3.8	13.2	5.7
Total	40.4	14.2	45.8	22.1	51.4	30.0

Public School Publishing Company. Specimen set 15¢; $2.50 per 100.

Reference: Williams, L. W. "Illinois Standardized Algebra Tests," *Journal of Educational Research*, 3:75–76, January, 1921.

Standard Diagnostic Tests for Elementary Algebra

H. R. Douglass (1921)

Series A, B; Forms I, II of each

The original Series A appeared at the date given above. A year or two later it was revised, Form II added, and both forms

of Series B prepared. The tests are especially intended for first-year students, but may also be used during the third semester of algebra. Series A consists of four subtests dealing, respectively, with addition and subtraction, multiplication, division, and simple equations. Each consists of ten examples arranged in order of difficulty, and involving different combinations of steps. Series B is in several parts: the first contains Test I on fractions and II on factoring; the second, III on formulae and fractional equations, and IV on simultaneous equations; the third, V on graphs; the fourth, VI on square roots, exponents, and radicals, and VII on quadratic equations. Each test of this series contains only five exercises.

The subtests of Series A deals with the four processes most frequently named by fifty-nine college and university teachers of mathematics as fundamental in first-year high-school algebra. Series B is based upon the opinions of twenty-one mathematics teachers as to what topics should be covered in a supplementary series of tests. For both series, Douglass selected suitable examples from quite a number of algebra texts. Weights were determined for the exercises originally used in Series A, but, because of the high correlation between weighted and unweighted scores, were omitted when the revised tests were issued. All of Series A may be administered within an ordinary high-school period, whereas Series B requires almost two hours.

Douglass announced the function of the tests as being to measure power in the fundamental and certain supplementary processes used in high-school algebra, and to yield diagnostic scores. The first appears to be rather satisfactorily fulfilled, but not the second. There are too few examples of each subtype to render the tests diagnostically of great value. Also it appears that bright pupils are not adequately tested. The published evidence as to the reliability of the tests differs, but averages about as follows:

$$r = .75,\ \text{P.E.}_{\text{meas.}} = 1.5,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .06,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .33.$$

Correlations of .52 with the Illinois, .65 with Hotz B and .72 with Hotz A have been reported. The published norms for the tests are not based upon a large number of pupils, but since

the schools represented are well scattered over the country, they may be considered fairly satisfactory. They are as follows for the end of the year:

Series	*Subtest*							*Total*
	I	*II*	*III*	*IV*	*V*	*VI*	*VII*	
A	7.8	7.1	6.5	7.3	—	—	—	29
B	2.4	4.1	3.1	3.6	2.5	2.7	3.4	22

Bureau of Administrative Research. Series A, 2¢ per copy, $1.60 per 100; Series B, 4¢ per copy, $3.50 per 100; scoring keys and class record sheets 3¢ each, free with 100 copies.

References: Douglass, H. R. "A Series of Standardized Diagnostic Tests for the Fundamentals of Elementary Algebra," *Journal of Educational Research*, 4:396–403, December, 1921.

———. "The Derivation and Standardization of a Series of Diagnostic Tests for the Fundamentals of First Year Algebra," *University of Oregon Publication*, Vol. 1, No. 8. Eugene, Oregon: University Press, 1921. 48 p.

———. "The Douglass Standard Diagnostic Tests for Measuring Achievement in First Year Algebra," *University of Oregon Publication*, Vol. 2, No. 5. Eugene, Oregon: University Press, 1924. 41 p.

Standard Survey Test for Elementary Algebra

H. R. Douglass (1928)

Tests I, II; Forms A, B of each

Each test consists of twenty-five exercises ranging from very easy addition to written problems. The tests are intended to conform to the recommendations of the National Committee on Mathematical Requirements, and to emphasize the solution of such equations, formulae, and graphs as are likely to be of greatest use both in later mathematics courses and outside of school. As is indicated by their title, these tests are intended for survey rather than diagnostic purposes. Test I covers the work of the first semester of high-school algebra, and Test II that of the second, but it is recommended that for testing at the end of the year both be used. The time for each is forty minutes. Douglass gives the following data on reliability:

	r	$P.E._{meas.}$	$\frac{P.E._{meas.}}{M}$	$\frac{P.E._{meas.}}{\sigma}$
Test I	.78	1.5	.07	.33
Test II	.74	1.1	.12	.35
Tests I and II	.86	1.7	.06	.25

Tentative norms for the end of the first and second semester are given as follows:

Test	*First quartile*	*Median*	*Third quartile*
I	16	20	24
II	6	9	11

Bureau of Administrative Research. 2¼¢ per copy, $2.00 per 100; manual and key, 3¢ each, free with 100 copies.

Columbia Research Bureau Algebra Test

A. S. Otis, J. B. Orleans, J. S. Orleans, and B. D. Wood (1927)

Tests 1, 2[4]; Forms A, B of each

Test 1 covers the first semester's work and Test 2 that of the second. Each consists of two parts, the first dealing with the mechanics of algebra, chiefly the solution of equations, and the second containing written problems. In the latter pupils are required to give the equations, as well as the values of the unknowns. Test 1 has thirty-five examples in Part I and eighteen problems in Part II; Test 2 has twenty in each. The actual working time is eighty minutes for Test 1 and one hundred for Test 2.

The purpose of these tests is announced as being "to provide a final examination or college entrance test in first year algebra . . . also to improve upon the previous tests of the kind by eliminating certain operations . . . which are of relatively minor importance and by providing *real* problems for solution." The reliabilities of the two tests are about as follows:

[4] Test 2 was originally published alone without any distinguishing number.

Test 1:

$$r = .92,\ \text{P.E.}_{\text{meas.}} = 3,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .10,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .20.$$

Test 2:

$$r = .90,\ \text{P.E.}_{\text{meas.}} = 2,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .08,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .22.$$

Norms for a few hundred cases on each test are given below.

Test	Semester	*Percentile*						
		5	*10*	*25*	*50*	*75*	*90*	*95*
1	I	11	15	23	33	42	48	51
2	II	8	10	13	17	23	28	33
2	IV	20	22	25	30	34	40	43

Especially on Test 2 the scores made on the two parts indicate that the reasoning problems are much more difficult than the examples.

World Book Company. Specimen set 30¢; Test 1, $1.20 per 25; Test 2, $1.30 per 25.

First Year Algebra Tests

Neva Carman (1929)

Tests 1, 2, 3, 4, 5, 6

Each of these tests is intended to cover the work of a six weeks period. It consists of four parts in various forms: completion, single-answer, multiple-answer, matching, and so forth. The total number of elements in the different tests varies, ranging from approximately sixty to one hundred. The time for each is forty minutes.

Harlow Publishing Company. Sample set 10¢; 75¢ per 25, $2.50 per 100.

Exercises and Tests in Algebra through Quadratics

D. E. Smith, W. D. Reeve and E. L. Morss (1926)

Test 1–224

This is probably the most elaborate and complete series of practice exercises or tests for algebra now available. The tests,

which cover very thoroughly the subject of algebra through quadratics, are grouped under thirteen topics, with from two to thirty-six on each. The number of exercises in the tests averages about thirty, each one consisting of as many exercises of the type included as can be conveniently placed on a single page. Varying time limits are given, most of which range from three to thirty minutes. Brief suggestions to pupils and also to teachers for their general use accompany the series. Apparently it is not intended that pupils score their work, but rather that they check the scoring after some one else has done it to ascertain that it is correct. The tests are in pairs, there being two dealing with each subtopic. It is perhaps needless to say that in the ordinary school with a total term of not to exceed two hundred days, and usually less than that number, it would be practically impossible to employ all of these tests. No norms or standards are suggested, and probably never will be.

Ginn and Company. Booklet containing all tests, 60¢.

Instructional Tests in Algebra, with Goals for Pupils of Varying Ability

Raleigh Schorling, J. R. Clark and Selma Lindell

(1927)

Tests 1–52

This series of tests, intended to cover at least the first year's work in algebra, is arranged in the approximate order in which the topics or processes are usually taught. Each contains from a few up to ninety-three examples. Ordinary written problems in algebra are not included. These are similar in purpose to those of Smith, Reeve and Morss rather than to regular standardized tests. The brief general directions suggest that the booklets be placed in the hands of the pupils about the sixth or seventh week of the year as supplementary material, and that they be inspected from time to time by the teacher. Pupils are to score their own work by reference to the answers. The same time, eight minutes, is allowed for each test. No space is given for pupils' work, since it is expected that this will be done on

blank paper. The space for recording scores at the end of each test provides for five trials. Three goals, denominated excellent, good, and passing, are given for each test. In general the first requires the giving of correct answers to from 90 to 100 per cent of the exercises, that labelled good to from 80 to 85 per cent, and that labelled passing to from 60 to 75 per cent. Though intended primarily for drill or instructional purposes, the large number of exercises of each type render these tests very useful for diagnosing and measuring ability in different processes.

World Book Company. Booklet containing all tests, answers, and directions, 28¢.

Diagnostic Tests and Drills in First Course Algebra

W. W. Hart (1929)

Tests 1–93

This series is similar to two just described in that it is not standardized but rather intended for practice or instructional purposes. Each test consists of a page of exercises, the number varying from six to eighty-five, depending upon the space needed for each. The first test deals with difficult addition combinations, the second with difficult subtraction combinations, and so on through a few other phases of arithmetic, simple formulae and equations, on up to quadratics, numerical trigonometry, functional relations, and so forth. On the back of each test is remedial instruction and drill material to accompany it. It is suggested that this may either be required on the part of pupils who make low scores on the tests or used as a second test later. The time limits of the tests vary, five minutes being most usual, with ten and fifteen, however, also frequent.

The tests are particularly intended to be used in connection with Wells and Hart's *First Year Algebras*, but may well be employed with most, if not all, texts in the subject.

D. C. Heath and Company. 56¢ per set.

Algebra Prognosis Test

J. B. and J. S. Orleans (1928)

Form A

This consists of an arithmetic test, followed by twelve subtests dealing with various phases of algebra. Each of the first ten subtests, and also Nos. 11 and 12 combined, is preceded by a lesson that explains what is to be done in the test immediately following. Thus the first lesson explains the algebraic use of letters to stand for numbers and the omission of the times sign between numbers and letters. Subtest 1, which follows, is composed of exercises calling for the substitution of numbers and letters, and the evaluation of such terms as $5a$, kn, $10xy$, and so forth. The average number of exercises in the subtests is about twelve, and the average length of the lessons about half a page. Approximately an hour and a half is required to give the test. It is suggested that the first portion extending through Subtest 7 be given within one high-school period, and the remainder within a second. The scoring of several of the subtests is somewhat complicated, but not extremely so.

As may be inferred from its name, the purpose of the test is to predict the probable success pupils will have in studying algebra. Results from several hundred pupils gave a correlation of scores on this test with achievement-test scores at the end of one semester of algebra of .71, and of scores on this test combined with ratings on the New York Rating Scale for School Habits,[*] of .82. No norms in the ordinary sense of the word have been published. The chances that pupils making various scores on the test have to succeed in a course in algebra are the same as for the Orleans-Solomon Latin Prognosis Test on p. 180.

World Book Company. Specimen set 20¢; $1.50 per 25.

II. Geometry

Geometry is one of several high-school subjects in which there has been an unusual amount of activity among test makers in the last few years. Up until about 1924 little had been done

[*] See page 428.

to follow up the start made a number of years earlier, but since that time tests have appeared at an average of several per year. Most of this activity has been directed to measuring pupils' ability in demonstrative rather than in intuitive geometry. Also almost all of it has been in plane rather than solid geometry. Although this is to be expected since many high schools do not offer the latter at all, and in practically all schools many more pupils take the former, yet it seems out of proportion that there should be perhaps six times as many standardized tests in the former as in the latter. This condition is partly remedied by the fact that the very few tests in solid geometry are of distinctly high merit, ranking much above the average of those in plane geometry.

Geometry Tests
J. H. Minnick (1918)
Tests A, B, C, D[6]

Although these tests are still available, they are mentioned here because they were the first widely used standardized tests in this subject rather than because of their merit as compared with other tests now available. In Test A five propositions are stated, for which the proper figures are to be constructed. In B there are four theorems accompanied by figures, and pupils are required to state what is given and what is to be proved. Test C gives four figures and certain facts about each, and asks pupils to state as many more facts as possible. In Test D figures, facts, and hypotheses are given for three exercises and pupils are to supply the proofs. Thirty minutes are allowed for each test. Minnick divided formal geometry into three chief parts and constructed these tests to deal with the usual steps in one of the three parts, the demonstration of theorems. Even for this limited purpose, however, they are not very satisfactory because the number of exercises is so small as not to constitute at all a sufficient sampling.

[6] In addition to these four, Test W, later dropped, was also prepared. It consisted of five exercises, each a partially completed drawing with a statement of what is given, and called for the making of any additional drawings necessary to prove certain facts.

Provision is made for two scores, a positive score based on the number of correct responses, and a negative score on the number of incorrect or unnecessary ones. This, together with the fact that there are several possible correct solutions for each exercise, renders scoring rather difficult, but gives results of greater diagnostic value than if a single score were used. Reliability figures for the single tests are:

$$r = .63,\ \text{P.E.}_{\text{meas.}} = 8,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .12,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .40.$$

For a rather small number of cases intercorrelations ranging from about .50 to .80 have been found between the various tests. Medians for the end of one year are:

Test	*A*	*B*	*C*	*D*
Positive	63	69	51	73
Negative	7.1	3.5	4.1	2.6

Public School Publishing Company. Sample set 20¢; $1.00 per 100.

References: Minnick, J. H. *An Investigation of Certain Abilities Fundamental to the Study of Geometry.* Philadelphia: University of Pennsylvania, 1918. 108 p.

———. "Certain Abilities Fundamental to the Study of Geometry," *Journal of Educational Psychology*, 9:83–90, February, 1918.

———. "A Scale for Measuring Pupils' Ability to Demonstrate Geometrical Theorems," *School Review*, 27:101–9, February, 1919.

Geometry Test

L. L. Thurstone (1919)

This test, which is no longer receiving much use, is briefly described along with the others of the series of Vocational Guidance Tests on p. 377.

Columbia Research Bureau Plane Geometry Test

H. E. Hawkes and B. D. Wood (1924)

Forms A, B

This, a revision of an earlier test by the same authors, deserves high rank among geometry tests. The first part contains

sixty-five true-false statements, and the second thirty-five numerical problems in geometry. Elements common to most of the texts used in the United States are included. The five books found in most texts are represented roughly by the following proportions of exercises: 4, 3, 3, 2, 1. Several times as many exercises as were finally included were tried out, and four equal forms of which only two have so far appeared prepared. The time is sixty minutes. Tryouts indicate higher correlations with secondary-school plane geometry marks than were found for traditional college-entrance examinations in this subject, the average correlations being about .55 and .50, respectively. Those with success in college mathematics were about .50 and .40. The reliability of Part I of the test is low, but that of Part II high, resulting in the following figures for the whole:

$$r = .93,\ \text{P.E.}_{\text{meas.}} = 3,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .08,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .18.$$

For over two thousand college freshmen and almost eight hundred Pennsylvania high-school seniors results are:

	Percentile						
	5	*10*	*25*	*50*	*75*	*90*	*95*
College freshmen	8	13	20	30	39	49	55
High-school seniors	6	9	15	23	31	39	45

In addition to the test proper there is a supplement to the manual of directions which may be used to augment the test proper. There are six statements on loci to be completed, four propositions of which the converses are to be given, four terms to be defined, and five propositions and constructions to be proved in the augmented material for each form. The time limits for these four extra parts are, respectively, fifteen, five,

Part	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
III	0	0	0	4	9	14	18
IV	0	0	3	9	10	13	14
V	0	0	4	5	9	13	15
VI	1	5	9	21	32	46	60

five and forty-five minutes. The following norms based on only five or six hundred college students are given for the several parts of the supplement:

World Book Company. Specimen set 25¢; $1.20 per 25.

Achievement Test in Plane Geometry
Raleigh Schorling and Vera Sanford (1925)
Forms A, B

This test, which is the result of several revisions of an earlier one by Schorling alone, contains five parts. The first is a completion test on geometrical facts; the second involves drawing conclusions from given data; the third judging the correctness of conclusions; the fourth analyzing constructions; and the fifth computation. There are sixty exercises in all, some of which call for more than one response. The time allowed for the actual work is fifty-two minutes. The material covered by this test falls within the limits defined by the National Committee on Mathematical Requirements and the terms and symbols employed are largely restricted to those found in the committee's report. The test is intended to be given at the end of the subject and is not suitable for use early in the year. The following data on reliability are perhaps slightly too low:

$$r = .70,\ \text{P.E.}_{\text{meas.}} = 6,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .18,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .36.$$

Correlations with teachers' judgments and with the results of the Regents Examinations tend to fall between .60 and .70. Norms from over twelve hundred cases are as follows:

Percentile	*5*	*10*	*25*	*50*	*75*	*90*	*95*
Norm	19	21	26	33	40	45	48

Bureau of Publications. Specimen set 35¢; 10¢ per copy; manual 25¢.

References: Sanford, Vera. "A New Type Final Geometry Examination," *Mathematics Teacher*, 18:22–36, January, 1925.
Symonds, P. M. *Ability Standards for Standardized Achievement Tests in*

the High School. New York: Bureau of Publications, Teachers College, Columbia University, 1927. 91 p.

Diagnostic Tests in Plane Geometry

O. W. Renfrow (1926)

Tests I, II; Forms A, B of each

The first test contains fifteen completion exercises dealing with definitions, axioms and postulates, nine construction and locus problems, six exercises in drawing the lines required to prove certain theorems, the completion of the proofs of six theorems, and nine exercises and problems, mostly numerical. In Test II are ten exercises on definitions, axioms, and postulates, ten construction and locus problems, five exercises requiring construction lines, five the completion of proofs, and ten exercises and problems. Test I covers the work of the first semester, and Test II that of the second. No time limit is set, but it is suggested that forty-five minutes should be sufficient for most pupils.

$$r = .88,\ \text{P.E.}_{\text{meas.}} = 1.5,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .03,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .23.$$

Standards for the end of the first and second semesters for Tests I and II, respectively, are announced as fifty-four and forty-three points.

Bureau of Administrative Research. Specimen set, single form 10¢, all forms 35¢; 4¢ per copy, 85¢ per 25.

Achievement Test in Plane Geometry

Maude McMindes (1926)

Forms A, B

Each form contains seventy-five true-false and thirty-one multiple-answer exercises, also thirty geometrical problems. Items dealing with principles contained in several of the texts and expressed in common terminology were selected after an analysis of a large number of geometry texts, although six well known ones were the chief basis. After a tryout of preliminary forms containing a large number of items, those included were

finally chosen. The working time is twenty minutes for Part I, ten for Part II, and thirty for Part III, and it is suggested that since the complete test is too long to be given in an ordinary high-school period, portions be given on each of two or three days.

This test appears to be among the two or three in this subject which represent the most careful selection and trying out of items included. It is claimed that the true-false and multiple-choice elements are unusually thought-provoking. The problem section is intended to minimize the mechanics involved while testing knowledge of geometric principles.

$$r = .77,\ \text{P.E.}_{\text{meas.}} = 5,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .13,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .33.$$

Norms based on more than a thousand responses to each form for the end of a year's study of geometry are announced as follows for Form A:

Percentile	*5*	*10*	*25*	*50*	*75*	*90*	*95*
Norm	16	19	27	38	49	60	67

Those for Form B tend to be slightly lower, the difference being about one point.

Public School Publishing Company. Sample set 15¢; 75¢ per 25.

Reference: O'Brien, F. P., and McMindes, Maude. "A Measure of Achievement in Plane Geometry," *University of Kansas Bulletin of Education*, Vol. 2, No. 3. Lawrence: Bureau of School Service and Research, University of Kansas, 1929, p. 16–18.

Geometry Tests

P. E. Webb (1926)

Form A

This contains five parts. In the first are eleven completion statements dealing with a given figure, in the second nine completion statements concerning geometrical facts, in the third

six ordinary multiple-answer exercises, in the fourth five multiple-answer exercises with one incorrect answer each, and in the fifth five different constructions to be made. A total of forty minutes' working time is allowed, divided among the various parts. The coefficient of reliability is .87. Norms for over one thousand pupils at the end of a year's study are:

Part	*I*	*II*	*III*	*IV*	*V*	*Total*
Median	16	6	4	3	16	39

Public School Publishing Company. Sample set 15¢; 75¢ per 25.

Geometry Tests

W. W. Hart (1927)

Tests I, II, III, IV, V

This series of tests is intended to accompany Wells and Hart's *Modern Plane Geometry*. Each is designed to be given at the completion of the corresponding book of geometry. Test I contains twenty-six figures for each of which from one to four questions are to be answered. The other tests contain respectively twenty-three, twenty-three, ten, and eleven similar exercises, except that in some of the exercises of Test III and in all those of Test V there are either no figures or the pupils are to draw them. Although the tests are intended to accompany a single text, it appears that they might fairly well be used in connection with others, since the points dealt with are generally included in geometry texts. The directions state that pupils should be allowed forty-five, thirty-five, forty, thirty and twenty minutes, respectively, if necessary. The following median scores are given:

Test	*I*	*II*	*III*	*IV*	*V*
Score	42	25	25	13	13

D. C. Heath and Company. $1.68 per 20 of whole series.

Geometry Tests

R. D. Perry (1928)

Tests I, II, III, IV, V

In this series as in some others the five tests are intended to accompany the common division of plane geometry into five books. Each consists of four subtests, the first containing true-false exercises dealing with definitions and theorems, the second computation exercises, the third multiple-answer exercises dealing with logical conclusions, and the fourth calling for geometrical constructions. The number of items in each subtest and also the time allowed varies somewhat, but the total working time for each test is the same, thirty-five minutes. Each test is intended to be given after completing the book or chapter upon which it is based and to facilitate so doing the teacher's key contains a table of some twenty textbooks indicating after what portion of each book each test should be given. The coefficient of reliability is announced as being about .89 for each of the tests. The following mean scores are given, based on about five thousand pupils in midwestern high schools:

Part	Test				
	I	*II*	*III*	*IV*	*V*
1	16	13	14	14	15
2	16	12	16	12	15
3	7	6	5	5	5
4	5	12	11	13	11
Total	44	43	45	43	46

Lafayette Printing Company. Specimen set 25¢; 75¢ per 25.

Plane Geometry Achievement Test

J. B. and J. S. Orleans (1929)

Tests 1, 2; Forms A, B of each

Test 1 is intended to cover the first half year's work and Test 2 that of the second half year. Each consists of five subtests. Those in the first deal with reasons, computation, completing proofs, diagrams, and proofs, and those in the second with the

same phases except that instead of reasons, loci are substituted. Part of the exercises are in multiple-answer form, part involve completion, part are similar to ordinary classroom work in geometry in that they require complete computation, proof, and so forth. Test 1 calls for forty-three detailed responses and twelve longer ones, and Test 2 for thirty-six and fifteen, respectively. Each requires somewhat more than an hour of working time. Average reliability data for single semester groups for the two tests are about as follows:

Test 1:

$$r = .85,\ \text{P.E.}_{\text{meas.}} = 5,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .08,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .26.$$

Test 2:

$$r = .71,\ \text{P.E.}_{\text{meas.}} = 9,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .15,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .36.$$

Correlations with school marks for ten different teachers vary from .66 to .88. Percentile norms based on about thirty-five hundred cases are as follows:

Test	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
1	30	37	48	61	75	88	95
2	24	29	42	58	76	90	95

Approximate medians for the separate parts are also given:

Test	Form	*Part*				
		I	*II*	*III*	*IV*	*V*
1	A	10	11	10	12	18.5
	B	10.5	10.5	9	10.5	21
2	A	10	8	6.5	18	15
	B	12.5	9.5	5.5	14.5	15.5

World Book Company. $1.20 per 25.

American Council Solid Geometry Test

H. W. Raudenbush, L. P. Siceloff, and B. D. Wood (1928)

Forms A, B

The first part of the test contains ninety-two true-false statements dealing with geometrical information of many sorts, and the second twenty-eight numerical problems in solid geometry. It is intended for both high-school and college use. Sixty minutes of actual working time are required. The reliability of this test is low.

$$r = .72,\ \text{P.E.}_{\text{meas.}} = 7,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .17,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .35.$$

Form A is about five points easier than B. Norms for three hundred pupils on Form A at the end of the course are:

Percentile	*5*	*10*	*25*	*50*	*75*	*90*	*95*
Score	16	21	30	41	57	71	81

Those for 747 Pennsylvania seniors are:

Semester	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
II	1	2	5	14	32	58	69
I	1	2	5	11	22	34	41

World Book Company. Specimen set 20¢; $1.25 per 25.

Seattle Solid Geometry Test Series

M. E. Morgan, W. T. Wait, August Dvorak (1928)

Tests I, II, III, IV, V, VI, VII, VIII, IX; Scales A, B

Tests I and II deal with lines and planes; III with dihedral and polyhedral angles; IV, prisms and parallelopipeds—volumes; V, pyramids; VI, cylinders and cones; VII, spheres, cylinders and cones; VIII, spherical polygons and spherical areas; IX, spherical volumes and areas. Scale A is intended to serve as a final examination, and Scale B as a pre-

liminary examination. All are composed of true-false statements. Each test has from thirty-two to thirty-four; the scales have sixty-three each. A is composed of seven on each of the nine test units, not the same as any included in the tests, whereas a number of those in B are in the tests. Many statements were developed and for several years tried out in a Seattle high school. Two revisions were then given in about forty high schools, and finally the present form constructed. No time limits are provided, but it is stated that pupils should be able to finish a test in from twenty to twenty-five minutes, and a scale in from forty-five to fifty. For a small number of cases a coefficient of reliability of .78 has been obtained. The following norms for the tests are given:

Test	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
I	14	15	18	21	24	26	27
II	17	18	20	23	26	29	30
III	16	17	20	23	26	28	29
IV	15	16	18	20	23	25	26
V	16	17	19	22	25	27	28
VI	13	14	17	20	23	25	26
VII	16	18	21	24	27	29	30
VIII	16	17	20	23	26	28	29
IX	16	17	19	22	25	28	29
Total	163	166	180	196	212	232	239
Scale A or B	33	35	38	42	47	51	52

Public School Publishing Company. Sample set 30¢; $5.00 per 25 of each test and Scale A; 50¢ per 25 of either scale.

Exercises and Tests in Plane Geometry

D. E. Smith, W. D. Reeve, and E. L. Morss (1928)

Most of this series of 160 tests deal with particular topics in plane geometry. There are also a number of review tests and at the end six that provide a general survey of the year's work. The tests contain from one up to about twenty exercises or items each. These include problems, proofs, constructions, completion exercises, true-false statements, matching exercises, multiple-answer statements, and so forth. The tests are not timed, but

appear to require from fifteen up to thirty or forty minutes each.

These tests are not standardized, and are similar to the same authors' tests in algebra in that they are intended for practice or instructional purposes. They are announced as fitting all modern courses and this claim appears to be rather well justified. The directions and suggestion to both pupils and teachers are decidedly limited. Answers are not provided, but there are very few cases in which competent teachers of the subject should have any doubt as to what they are.

Ginn and Company. Booklet containing all tests, 48¢.

Unit-Achievement Tests in Plane Geometry

Ruth O. Lane and H. A. Greene (1929)

Tests 1, 2, 3, 4, 5, 6; Forms A, B of each

This series is intended to cover the plane geometry commonly taught in high school, especially measuring pupils' ability to follow through the steps of reflective thought until they arrive at valid conclusions and to think through the proofs of the main theorems. Test 1, on Fundamental Ideas of Geometry, consists of thirty-four elements; Test 2, Parallel Lines and Triangles, of sixty; Test 3, Rectilinear Figures, of fifty-nine; and the other three on The Circle, Proportion and Similar Polygons and Areas of Polygons, of forty-five each. The items included are arranged in the usual order so that the tests may be given approximately at the ends of six-week periods. In each test except the first there are four parts, containing exercises in recall, completion, multiple-response, and order of procedure form. The time varies from thirty-five to thirty-eight minutes of actual work.

These tests are the result of a number of years' experimental work, part of which was with this series. They deserve very high rank with regard to the study, care, and amount of experimental work that entered into their construction. An attempt has been made to insure high validity by testing on all theorems recommended by the National Committee, by using no word not common to eighteen widely used texts or which the National Committee advises against, by avoiding definitions not uniformly

agreed upon, and other similar precautions. Average reliability data for the six tests are as follows:

$$r = .85,\ \text{P.E.}_{\text{meas.}} = 2,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .08,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .26.$$

Medians based on five or six hundred pupils for each form of each test are as follows:

Part	*Test*					
	1	*2*	*3*	*4*	*5*	*6*
1	—	8	9	5	5	4
2	—	6	8	5	3	4
3	—	5	6	4	3	4
4	—	14	9	8	7	5
Total	20	33	32	22	18	17

Percentile norms for total scores are also given:

Test	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
1	10	12	16	20	24	27	29
2	14	17	23	33	38	44	47
3	14	17	25	32	37	40	43
4	10	12	15	22	27	30	32
5	8	9	12	18	24	30	32
6	7	8	12	17	22	28	30

Ginn and Company. Tests 1, 4, 5, 6, 68¢ per 25; Tests 2, 3, 80¢ per 25; manual 20¢.

Geometry Prognosis Test
J. B. and J. S. Orleans (1929)
Form A

In general arrangement this is similar to the Algebra Prognosis Test by the same authors. There are ten subtests calling for from eleven to twenty-one responses apiece. Each except the final summary test is preceded by a lesson dealing with the points covered in the test. Pupils are to study each lesson and then take the corresponding test. The items dealt with include

axioms, relationships of angles, understanding and analysis of geometric relationships and statements, bisection, notation, and problems. Seventy minutes of working time are required. Correlations ranging from .67 to .77 have been found between scores on this test and those on an objective achievement test at the end of the first semester. These resulted from the use of a preliminary form. Revision and extension have improved the test so that it is stated that the correlation between Form A and a satisfactory achievement test is at least .80. Norms of the ordinary kind are not published for this test. The expectation of success according to test scores is the same as for the Orleans-Solomon Latin Prognosis Test and may be found on page 180.

World Book Company. Specimen set 20¢; $1.70 per 25.

III. Trigonometry

Although trigonometry is taught in a considerable number of secondary schools and in practically all institutions of higher learning, there has been, in so far as the writer is aware, only one serious effort to construct a standardized test in this subject. Although this test, which will be described below, is relatively satisfactory, it would certainly be desirable to have several others covering this same subject. Perhaps one reason why more trigonometry tests have not been developed is the recent tendency to displace trigonometry by a composite course that draws its content from several branches of mathematics, of which trigonometry is only one. However, there has been little activity in developing tests for such a course as this so that trigonometry is not adequately taken care of either as a separate subject or in combination with others.

American Council Trigonometry Test

H. W. Raudenbush, L. P. Siceloff, and B. D. Wood

(1928)

Forms A, B

This, the only standardized test now available in this subject, is divided into four parts. These contain, respectively,

thirty-seven multiple-answer exercises, twenty-four true-false statements, twenty-eight single-answer statements, and twenty-two more of the same sort. In the last two the answers are all numerical. A total of seventy minutes' working time is required. This, in common with the other American Council Tests, deserves high rank as a test. A tentative median of thirty-five points for public high-school students and fifty points for private secondary-school students has been announced. Norms for over twenty-five hundred Pennsylvania high school seniors are:

Semester	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
II	10	14	21	31	44	62	68
I	8	14	19	27	36	46	52

World Book Company. Specimen set 20¢; $1.25 per 25.

IV. General Mathematics

Although, as was stated in the last section, so-called general, unified or composite courses in mathematics have increased in number during the past decade or so very few tests suitable for use therein have been published. Such courses are perhaps most commonly offered in the junior high school, but they may also be found in the senior high school and in the freshman or rarely even freshman and sophomore years of college. Since the chief content of such courses in Grades VII, VIII, and IX is arithmetic, and since this is an elementary rather than a high-school subject, no mention has been made of tests in general mathematics designed primarily for these grades. Of the few efforts along this line looking to senior high-school tests there is only one that appears to deserve mention in this section.

In addition to attempts to predict the likelihood of success in algebra and geometry specifically, such as are reported by the Orleans tests described under those subjects, other efforts have been made to measure a somewhat more general type of mathematical capacity or likelihood to succeed in mathematics courses. Most of these have been by means of combining already existing tests of several sorts, frequently including arithmetic.

In very few instances have any new or distinctive tests been developed. The only such tests that seem to the writer to deserve mention are the one described immediately below and the mathematics tests in the Iowa Placement Series.

Test of Mathematical Ability
Agnes L. Rogers (1918)

Although this is one of the best known series of tests in the field of mathematics, it has little place in the work of the ordinary teacher. It consists of six tests or pairs of tests as follows: geometry, algebraic computation, interpolation in number series, superposition, language scales, mixed relations or analogies. After considerable critical and experimental work Miss Rogers selected this combination as best suited to predict mathematical ability. Evidently, however, the test cannot be used until after pupils have studied a considerable amount of algebra and geometry. The actual working time for the complete test is sixty minutes, but almost twenty-five more are required for the necessary directions and explanations. The average coefficient of reliability of the test appears to be at least .80. The intercorrelations of the parts tend to run from about .30 to .50, although there are some both higher and lower. The correlation with future mathematical achievement may be expected to be from .60 to .80.

As is true of most prognostic tests, norms of the ordinary sort are not generally used. The following scores for small groups of pupils are available, however:

Semesters of study	*Percentile*				
	10	*25*	*50*	*75*	*90*
Algebra 2 + Geometry 1	270	328	398	477	525
Algebra 2 + Geometry 1 *	408	470	524	584	664
Algebra 2	275	360	444	500	625
Algebra 1 or less	181	250	326	414	486

* This differed from the first group in being very highly selected and in that the geometry studied was demonstrative.

It is suggested that pupils who score above 650 points on the tests are very superior and should be allowed to progress at

their own rate; that those with scores above 550 should cover the work more rapidly than the ordinary high-school class; those between 350 and 550 should be in a regular class; those between 250 and 350 can probably cover two years' work in three; and those below 250 are so inferior that it is doubtful if they should carry further work in mathematics.

On the whole it appears that the Orleans Algebra and Geometry Prognosis Tests are to be preferred to this series.

Bureau of Publications. Specimen set 60¢; 10¢ per copy, $7.00 per 100; manual and scoring stencils 50¢.

References: Rogers, A. L. "Experimental Tests of Mathematical Ability and Their Prognostic Value," *Teachers College, Columbia University, Contributions to Education*, No. 89. New York: Bureau of Publications, Teachers College, Columbia University, 1918, 118 p.

Mensenkamp, L. E. "Tests of Mathematical Ability and their Prognostic Values—A Discussion of the Rogers Tests," *School Science and Mathematics*, 21:150–62, February, 1921. Also in: *High School Conference Proceedings*, 1920. Urbana: University of Illinois, 1921, p. 201–10.

Iowa Placement Examinations

MATHEMATICS APTITUDE AND TRAINING

G. D. Stoddard, E. W. Chittenden, C. E. Seashore, and G. M. Ruch (1924)

Forms A, B

These tests are described along with the other Iowa Placement Examinations on page 380. They, in common with the rest of this series, are of high merit.

Mathematical Achievement Tests, Senior Mathematics

E. R. Breslich (1928)

Tests I, II, III, IV, V, VI, VII, VIII, IX, X, XI

These tests are intended to cover the work of the tenth grade in mathematics as selected and organized by their author. They cover the eleven chapters of his *Senior Mathematics, Book II.* Each consists of a four-page folder containing a number of various types of objective or near objective exercises dealing with a particular topic. By far the greatest part of the content covered

is geometry, though there is some algebra and a very little trigonometry.[1] The number of responses called for by each test varies, the greatest being less than fifty. The topics covered by the tests are as follows:

I. Lines and angles
II. Parallel and perpendicular lines
III. Quadrilateral lines
IV. Proportional line segments
V. Loci
VI. Similar figures, theorem of Pythagoras
VII. Numerical trigonometry
VIII. The circle
IX. Measurement of angles by arcs, proportional segments in the circle
X. Regular polygons
XI. Areas

No time limits are given as pupils are to be allowed all the time needed. Also no norms are given, and it is stated that none are needed "because all of the items of each test are essential." It appears that most pupils would finish any one of the tests within an ordinary high-school period, but probably some of the slower ones would require considerably more time.

University of Chicago Press. 75¢ per 25.

BIBLIOGRAPHY

I. Algebra

Cawl, F. R. "Practical Uses of an Algebra Standard Scale," *School and Society*, 10:88–90, July 19, 1919.

Childs, H. G. "The Measurement of Achievement in Algebra," *Third Indiana University Conference on Educational Measurements*. Bloomington: Extension Division, Indiana University, 1917, p. 171–83.

Dalman, M. A. "Hurdles, A Series of Calibrated Objective Tests in First Year Algebra," *Journal of Educational Research*, 1:47–62, January, 1920.

Deam, T. M. "Diagnostic Algebra Tests and Remedial Measures," *School Review*, 31:376–79, May, 1923.

Harris, Eleanora. *A Study of the Hotz First-Year Algebra Scales and the Rugg-Clark Standard Algebra Tests*. Chicago: University of Chicago Press, 1919. 97 p.

Harris, Eleanora and Breed, F. S. "Comparative Validity of the Hotz Scales

[1] The exercises in the field of trigonometry deal with material also frequently taught in geometry or algebra or both.

. and the Rugg-Clark Tests in Algebra," *Journal of Educational Research*, 6:393–411, December, 1922.

Monroe, W. S. "Measurements of Certain Algebraical Abilities," *School and Society*, 1:393–95, March 13, 1915.

———— "A test of the Attainment of First-Year High-School Students in Algebra," *School Review*, 23:159–71, March, 1915.

Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction*. Yonkers, New York: World Book Company, 1927, p. 72–82.

Rugg, H. O. "The Experimental Determination of Standards in First-Year Algebra," *School Review*, 24:37–66, January, 1916.

Rugg, H. O. and Clark, J. R. "The Improvement of Ability in the Use of the Formal Operations of Algebra by Means of Formal Practice Exercises," *School Review*, 25:546–54, October, 1917.

————. "Scientific Method in the Reconstruction of Ninth-Grade Mathematics," *Supplementary Educational Monographs*, No. 7. Chicago: University of Chicago Press, 1918. 189 p.

————. "Standardized Tests and the Improvement of Teaching in First-Year Algebra," *School Review*, 25:113–32, 196–213; February, March, 1917.

Schmitz, Sylvester. "Measurements in First-Year Algebra," *Catholic University, Educational Research Bulletin*, Vol. 1, No. 9. Washington, D. C.: Catholic Education Press, 1926. 40 p.

Smith, D. E. "On Improving Algebra Tests," *Teachers College Record*, 24: 87–94, March, 1923.

Symonds, P. M. *Measurement in Secondary Education*. New York: The Macmillan Company, 1927, p. 113–25.

————. "Special Disability in Algebra," *Teachers College, Columbia University, Contributions to Education*, No. 132. New York: Bureau of Publications, Teachers College, Columbia University, 1923. 89 p.

Thorndike, E. L. "An Experiment in Grading Problems in Algebra," *Mathematics Teacher*, 6:123–34, March, 1914.

Thorndike, E. L., et al. *The Psychology of Algebra*. New York: The Macmillan Company, 1923. 483 p.

Upton, C. B. "Tests in Algebra," *The Reorganization of Mathematics in Secondary Education*. Mathematical Association of America, 1923, p. 323–70. (Obtainable from J. W. Young, Dartmouth College, Hanover, New Hampshire.)

II. Geometry

Haertter, L. D. "Use of the Inventory Test in Plane Geometry," *Mathematics Teacher*, 19:147–54, March, 1926.

Irwin, H. N. "A Preliminary Attempt to Devise a Test of the Ability of High-School Pupils in the Mental Manipulation of Space Relations," *School Review*, 26:600–5, 654–70, 759–72; October, November, December, 1918.

Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction*. Yonkers, New York: World Book Company, 1927, p. 84–91.

Saunders, M. O. "Geometry Test," *Fourth Yearbook, Chicago Principals'*

Club. Chicago: Chicago Principals' Club, 1929, p. 156–59.

Stockard, L. V., and Bell, J. C. "A Preliminary Study of the Measurement of Abilities in Geometry," *Journal of Educational Psychology*, 7:567–80, December, 1916.

Symonds, P. M. *Measurement in Secondary Education.* New York: The Macmillan Company, 1927, p. 125–28.

Upton, C. B. "Tests in Geometry," *The Reorganization of Mathematics in Secondary Education.* Mathematical Association of America, 1923, p. 376–95. (Obtainable from J. W. Young, Dartmouth College, Hanover, New Hampshire.)

IV. General Mathematics

Breslich, E. R. "Testing as a Means of Improving the Teaching of High School Mathematics," *Mathematics Teacher*, 14: 276–91, May, 1921.

Courtis, S. A. "The Measurement of High School Mathematics," *School Science and Mathematics*, 18:507–26, June, 1918.

Kelley, T. L. "Values in High School Algebra, and Their Measurement," *Teachers College Record*, 21:246–90, May, 1920.

Langlie, T. A. "A Comparison of 'Aptitude' and 'Training' Tests for Prognosis," *Journal of Educational Psychology*, 19:658–65, December, 1928.

Reeve, W. D. *A Diagnostic Study of the Teaching Problems in High School Mathematics.* Boston: Ginn and Company, 1926. 117 p.

————. "The Place of New-Type Tests in Teaching Mathematics," *Teachers College Record*, 29:693–703, May, 1928.

Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction.* Yonkers, New York: World Book Company, 1927, p. 70–72, 82–84, 91–92.

Young, J. W. (Chairman). *The Reorganization of Mathematics in Secondary Education.* Mathematical Association of America, 1923. 652 p. (Obtainable from J. W. Young, Dartmouth College, Hanover, New Hampshire.)

CHAPTER VIII

SCIENCE

Introduction.—Although the various subjects classed under high-school science are not provided with as many tests as exist in Latin, mathematics, and some other fields of secondary instruction, yet each of the four sciences most commonly taught in high school is represented by a few tests of sufficient merit to be described in this chapter. Not much was done toward the construction of science tests until after several years' activity in a number of other secondary subjects; therefore most of the existing tests are fairly new and fewer of them are of little value because they do not conform to present curricula than is true in many other subjects.

Several difficulties exist of which the maker of science tests must take account. Not only is there at least as great lack of agreement as to the content of high-school science courses as exists in regard to other high-school subjects, but there is also considerable variation among schools in the sciences actually offered. Still further there is no unanimity or anything approaching it as to the year in which particular sciences are offered nor as to the order in which they are studied if more than one is taken. Moreover there is commonly only one year of each science offered in high school so that there is much less opportunity for a continuous testing program than is true in the case of subjects which continue through two or more years.

As is true in most other subjects also the science tests constructed to date place most emphasis on the measurement of the knowledge acquired by pupils rather than on such other desirable outcomes at attitudes, interest, laboratory skill, and so forth. A few tests make use of diagrams, charts, and pictures in various ways, ordinarily to have parts identified, but this tendency has not been carried nearly so far as appears desirable. Certainly in general science, biology, and physics, although perhaps to a lesser

degree in chemistry, this method of testing has considerable value. Practically nothing has been done along the line of testing laboratory ability. It would be possible to set up standard laboratory conditions in which pupils should be placed with given instructions, equipment, and so forth, and measure how rapidly and well they perform the tasks indicated. It is certainly to be hoped that something of this sort will be done within the near future.

I. General Science

Although general science is frequently taught in the eighth and sometimes even in the seventh grade, especially when these grades are organized as part of a junior high school, it is commonly regarded as a high-school subject. The chief reason for this is that it is more commonly offered in the ninth grade than any other and sometimes even above this. Probably because it is relatively new as a subject there is less agreement as to the content of general science courses than is true of the other sciences commonly taught in high school. The general tendency illustrated both by text-books and tests is to place major emphasis upon subject matter most closely related to physics and biology with less upon that in the fields of chemistry, the earth sciences, and so forth. There is, however, such great variation between the work offered in different schools that anyone who is choosing a test in this subject should examine it critically from the standpoint of how well it covers the course as actually taught and also of how many of the test items are included in the course.

General Science Test

G. M. Ruch and H. F. Popenoe (1923)

Forms A, B

This test is more or less a revision of an earlier test constructed by Ruch which, however, never reached the stage of satisfactory completion and standardization. It has two parts of which the first contains fifty multiple-answer exercises, unusual in that there are seven suggested answers, and the second twenty diagrams with several completion statements about each to be filled in. The earlier test was based on an analysis of the content

of all the general science texts and manuals available in 1919. For the present test several additional texts published since 1919 were examined. Slightly more than one-third of the items deal with physics and mechanics, slightly less than one-third with biology, about one-fifth with earth sciences; and one-eighth with chemistry. The actual working time for the two parts is forty minutes, within which about 90 per cent of pupils ordinarily attempt every item.

$$r = .83,\ \text{P.E.}_{\text{meas.}} = 4,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .12,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .28.$$

Percentile norms for the end of a semester and of a year based on a total of about one thousand cases are given below.

Percentile	*10*	*25*	*50*	*75*	*90*
Year	23	28	36	45	54
Semester	20	24	28	36	43

Although this is one of the first standard tests constructed in general science, it still ranks among the best.

World Book Company. Specimen set 20¢; $1.30 per 25.

References: Ruch, G. M. "A New Test in General Science," *General Science Quarterly*, 7:188–96, March, 1923.

———— "A Range of Information Test in General Science," *General Science Quarterly*, 4:257–62, November, 1919.

———— "Range of Information Test in General Science: Preliminary Data on Standards," *General Science Quarterly*, 5:15–19, November, 1920.

Symonds, P. M. *Ability Standards for Standardized Achievement Tests in the High School*. New York: Bureau of Publications, Teachers College, Columbia University, 1927. 91 p.

General Science Scales

August Dvorak (1924)

Forms R–1, S–2, T–2

The first of these forms is intended for use fairly early in the year to provide a basis for diagnosis and classification, whereas the other two are equivalent forms to be used later. Each consists of sixty multiple-answer exercises arranged in order of increasing difficulty. From an examination of compilations of gen-

eral science material, attempts at measurement, and textbooks, Dvorak accumulated about six hundred test items. Three hundred of these were chosen, criticized by teachers, and tried out on more than ten thousand pupils, and on the basis of these results the 180 that compose the three scales were selected. About three-sevenths of these pertain to biology, three-eighths to physics, one-eighth to earth sciences and one-tenth to chemistry. It is suggested that twenty minutes is about the right time allowance, but that sufficient time be given that all pupils attempt all the items. Provision is made for computing scores in terms of difficulty values which are themselves in terms of the probable error. The various data presented as to reliability do not harmonize. Apparently the best figures are:

$$r = .85,\ \text{P.E.}_{\text{meas.}} = 4,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .05,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .26.^{1}$$

Correlations averaging slightly above .70 with general science marks have been found. Norms for seven hundred high-school freshmen at the end of a year of general science are given as follows:

Percentile	*5*	*10*	*25*	*50*	*75*	*90*	*95*
Norm	56	61	70	80	90	99	104

Public School Publishing Company. Sample set 20¢; 50¢. per 25.

References: Dvorak, August. "A Study of Achievement and Subject Matter in General Science," *General Science Quarterly*, 10:289–310, 367–96, 445–74, 525–42; November, 1925; January, March, May, 1926.

——— *A Study of Achievement and Subject Matter in General Science*. Bloomington, Illinois: Public School Publishing Company, 1926. 100 p.

[1] The manual of directions states that the probable error of estimate is not greater than two points, which would mean that the probable error of measurement was somewhat smaller. However, this does not seem to agree with the available data as to the values of the coefficient of reliability and the standard deviation.

General Science Examination

J. T. Giles, S. M. Thomas, and H. M. Schmidt (1924)

Series A, Numbers 1–22; B, Numbers 1–24

Although these tests have never been satisfactorily standardized, they appear to deserve mention because of their comprehensiveness. Series A is composed of eleven pairs, and B of twelve, Numbers 1 and 2, 3 and 4, and so forth being paired, each dealing with one of the topics into which general science was divided. The first test of each pair consists of fifty true-false statements, and the second of fifty multiple-answer statements. Thus the complete series provides well over a thousand items covering the year's work.

Eauclaire Book and Stationery Company. Single set of either series 30¢, $3.00 per 12.

General Science Test

S. R. Powers (1926)

Forms A, B

The three parts of this test contain a total of seventy multiple-answer exercises in somewhat different forms intended to cover a year course in general science. After examining a number of texts and courses of study, Powers reduced the resulting one thousand items to two hundred on the basis of experimental try-outs and teachers' judgments. From these the items found in the tests were selected. About one-third of them are concerned with biology, as many with physics, one-sixth with earth sciences and one-eighth with chemistry. Thirty-five minutes is the time.

$$r = .85,\ \text{P.E.}_{\text{meas.}} = 4,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .08,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .26.$$

Test scores correlate from .60 to .70 with school marks, almost .70 with Regents Examination marks, and more than .50 with I.Q.'s. Norms for Form A based on between eight and nine hundred cases are as follows:

Percentile	*10*	*25*	*50*	*75*	*90*
Year	34	42	50	60	73
Semester	20	24	30	39	49

It appears that at the end of the first semester Form B scores run several points higher than those on Form A, but that for the end of the year there is no material difference. Also boys' scores average about five points higher than those of girls.

Bureau of Publications. Specimen set 10¢; $2.00 per 100.

References: Powers, S. R. "Objective Measurement in General Science," *Teachers College Record*, 29:345–49, January, 1928.
Pruitt, C. M. "Objective Measurement in General Science," *General Science Quarterly*, 12:517–24, May, 1928.

Tests on Everyday Problems in Science

C. J. Pieper and W. L. Beauchamp (1927)

Units I–XVII; Factual, Major Ideas Test on each; Forms A, B, of each

These are not a series of ordinary standardized tests, but are intended for diagnostic and measurement purposes in connection with *Everyday Problems in Science* by the same authors. The seventeen major objectives for a general science course were determined and a factual and a major ideas test constructed over each. Each factual test calls for fifty responses to direct questions, completion statements, multiple-answer and alternative statements, and each major ideas test contains fifteen to twenty-five incomplete statements to be completed by choosing the proper one of four suggested possibilities. Although the authors state definitely that these tests are not intended for use except in connection with the textbook named above, it appears that the material in at least some of them is commonly enough taught that they might well be used in general science courses where this text is not employed. There appear to be no definite time limits, but apparently practically all pupils should be able to complete each test within fifteen minutes or thereabouts.

Scott, Foresman and Company. 2¢ per copy; 20% discount on orders of $5.00 or more.

READING SCALES IN GENERAL SCIENCE
M. J. Van Wagenen (1921)
Forms A, B

A description of these scales will be found on page 143, along with that of the whole series of which they are a part.

II. Physics

The subject of physics appears to have been the first high-school science to interest test makers. This was perhaps due to the fact that the subject was comparatively well established and that although there was no agreement as to the content of the course, yet the amount of disagreement was less than in the case of some of the other sciences. Notwithstanding this earlier start, however, the number of worthwhile tests in this subject is no greater than in the three other fields of science dealt with in this chapter.

Although the tendency is not as pronounced in physics as in chemistry, some of the tests in this subject are suitable for use in college as well as in high school. Indeed there are even one or two designed specifically for the college or university and not appropriate to high-school use at all.[2]

PHYSICS TEST
L. L. Thurstone (1919)

This test, which is no longer receiving much use, is briefly described along with the others of the series of Vocational Guidance Tests on page 377.

IOWA PHYSICS TESTS
H. L. Camp (1921)
Series A, B, C; Forms 1, 2 of each

The three series named above deal with mechanics, heat and

[2] Probably the best series of this sort is that constructed by C. J. Lapp and known as the Iowa Achievement Examinations in College Physics. There are eight tests in the series, with two forms of each, and they cover the work of two semesters. They may be secured from the Bureau of Educational Research and Service at $3.00 per 100.

electricity, and magnetism, respectively. Each test consists of eleven or twelve exercises, many of which are numerical problems. Camp divided high-school physics into five main phases, the three just named and light and sound. On the basis of Starch's determination of 102 principles common to five commonly used high-school physics texts,[3] exercises were constructed in the various phases dealt with, and after trying them out, those included in the scales were selected as best for testing purposes.[4] The exercises in each test are arranged in scalar order, the most difficult one being not far from three times as hard as the least difficult. Forty-five minutes are allowed for Series A and B, and forty for Series C, so that time is not a major factor in determining the scores made. It will be readily seen that tests as short as these yield only general measures of ability in the fields dealt with.

Provision is made for computing scores by adding the values or weights. This is somewhat laborious, but in a test with as few exercises as this it cannot be assumed that the correlation between weighted and unweighted scores will be as high as if there were a fairly large number. Because of the form of exercise used, scoring is not entirely objective. Norms from Iowa high schools for the end of the year are given for boys and girls separately since large differences appear.

Percentile	*25*	*50*	*75*
Boys	25	40	53
Girls	18	31	44
Both	21	34	49

Public School Publishing Company. Sample set 15¢; 50¢ per 25.

References: Camp, H. L. "Scales for Measuring Results of Physics Teaching," *University of Iowa Studies in Education*, Vol. 2, No. 2. Iowa City: University of Iowa, 1921. 51 p.

———— "Scales for Measuring Results of Physics Teaching," *Journal of Educational Research*, 5:400–5, May, 1922.

[3] Starch, Daniel. *Educational Measurements*. New York: The Macmillan Company, 1916, p. 188.

[4] The first reference given below contains not only the exercises employed in the tests, but about three hundred sixty others, for each of which the difficulty value and average time was determined, with the correct answers.

Physics Scales

J. M. Hughes

Information R, Division I; Information S, Division II; Thought R, Division I; Thought S, Division II

Each scale consists of thirty exercises arranged in order of increasing difficulty. Most of those in the thought scales are numerical problems, the others are not. The distinction between the exercises in the thought and those in the information scales is not at all clear, since many of the former apparently require information only. One-third of each Division I scale contains easier exercises than are found in the Division II scales, and likewise one-third of each of the latter more difficult exercises than are found in the former. In a few cases the same exercises are used in both divisions. The items included were selected according to results from a questionnaire sent to physics teachers and textbook analysis. The working time is twenty minutes for the information scale and thirty-seven for the thought scale. Reliability data for the information tests are:

$$r = .85,\ \text{P.E.}_{\text{meas.}} = 3,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .04,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .26.$$

Those for the thought tests are slightly higher.

Norms, apparently for the end of one year of high-school physics, are $Q_1 = 80$, Md. $= 88$, $Q_3 = 96$.

Public School Publishing Company. Sample set 15¢; 50¢ per 25.

Columbia Research Bureau Physics Test

H. W. Farwell and B. D. Wood (1926)

Forms A, B

This test of general achievement in physics contains 144 true-false information and reasoning statements covering the five main divisions of first-year physics. About one-sixth of the statements are based upon problems stated in the test, whereas the majority are not. The test is stated to cover "the materials common to the most popular and widely used high-school and

elementary-college textbooks, and also the essentials of the requirements laid down by recognized college-entrance examining agencies." About one-third of the statements deal with electricity, about one-sixth each with mechanics, heat and light, about one-twelfth with sound, and about one-eighth with miscellaneous topics. The working time is seventy-five minutes.

$$r = .90,\ \text{P.E.}_{\text{meas.}} = 5,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .10,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .21.$$

High-school norms for the end of the year based on over fourteen thousand students taking the New York State Regents Examinations are as follows:

Percentile	*5*	*10*	*25*	*50*	*75*	*90*	*95*
Norm	12	18	31	46	62	79	90

It appears that Form B is two or three points more difficult than Form A.

World Book Company. Specimen set 25¢; $1.30 per 25.

Reference: Wood, B. D. "The Regents Experiment with New-Type Examinations in French, Spanish, German, and Physics, June, 1925," *New York Experiments with New-Type Modern Language Tests. Publications of the American and Canadian Committees on Modern Languages*, Vol. 1. New York: The Macmillan Company, 1927, Part 2.

Harvard Elementary Physics Test

N. H. Black and Frances M. Burlingame (1927)

Forms A, B

Part I of this test contains twenty-six true-false statements, Part II eleven completion statements, and Part III twenty-three multiple-answer exercises. A number of the exercises are evidently intended to measure reasoning or problem solving ability, but most of them deal merely with items of information. The topics dealt with were found to be common to the College Entrance Examination Board's definition of the requirements in physics, the Syllabus in Physics for Secondary Schools by the Regents of the University of the State of New York, and eight

of the most commonly used high-school physics textbooks. The items included were selected from 320 as being those which differentiated best among good, average, and poor students. The time is thirty-five minutes.

$$r = .80,\ \text{P.E.}_{\text{meas.}} = 2.5,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .08,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .30.$$

Norms based on over two thousand cases from schools in twenty-four states are $Q_1 = 27$, Md. $= 33$, $Q_3 = 40$.

Ginn and Company. $1.00 per 30.

Iowa Placement Examinations

Physics Aptitude and Training

G. D. Stoddard, C. J. Lapp, C. E. Seashore, and G. M. Ruch (1924)

Forms A, B

These tests are described on page 380 along with the others of the same series.

Michigan Instructional Tests in Physics

P. V. Sangren and W. G. Marburger (1929)

Initial Test; Tests 1–22; Final Examination, Form A

As is indicated above, this series consists of one test to be given at the beginning of physics instruction to determine pupils' preparation therefor, twenty-two that cover different topics and are to be given during the course, and one for the end. The Initial Test contains thirty single-answer exercises dealing with physics information, twenty-five problems dealing with arithmetic, algebra, geometry, and quantative relations, and three paragraphs from physics texts followed by a total of twenty questions to be answered after reading them. Each of the twenty-two tests covers the two sides of a single sheet and contains from twenty to forty test elements in either single-answer or multiple-answer form. In the final examination there are 120 single- and multiple-answer exercises of which a number are mathematical problems grouped under the five heads of mechanics, heat,

magnetism and electricity, sound, and light. There are no definite time limits for the tests, but they are intended to be given within class periods of forty-five minutes. The content of the twenty-two instructional tests was based upon an analysis of five most commonly used textbooks. The items called for in the Final Examination are a random sampling of those in the previous tests.

The coefficients of reliability of the twenty-two tests range from .46 to .91, averaging about .80. That of the Final Examination is .94. Correlations between the various tests of the series range from .13 to .75, averaging about .50. Correlations between final semester marks and separate test scores likewise average about .50. The median scores for about two hundred pupils on the Initial Test and seven or eight hundred on each of the others are as follows:

Test	*Initial*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Median	56	26	27	23	18	23	21	27	21	20	21	23

Test	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*Final*
Median	24	30	17	28	21	19	22	13	21	18	18	91

Public School Publishing Company. Complete sample set 60¢. Initial Test, sample set 10¢; 50¢ per 25. Instructional Tests, 30¢ per series. Final Examination, sample set 10¢; $1.00 per 25.

Instructional Tests in Physics

E. R. Glenn and E. S. Obourn (1930)

Tests 1–25

This series of tests is the result of ten years of experimentation with four experimental editions. Each test covers two or three pages of a booklet and contains from about twenty up to fifty elements dealing with one or several topics. For example, Test 2 is on liquid pressure and Pascal's Law, Test 6 on energy, force, and motion, and Test 13 on static electrical phenomena. There are no time limits stated, but pupils are expected to record the time required on each test. Apparently none of the

tests should easily be completed within less than an ordinary class period.

The coefficients of reliability of the various tests range from .70 up to .88, the probable errors of measurement from 2 to 5, their ratios to the means from .09 to .25, and to the standard deviations from .23 to .37. Percentile norms are given in the manual, but, because of the space that would be required to do so, are not reproduced here. The medians are as follows:

Test	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Median	13	14	17	18	22	28	21	16	23	26	18	16

Test	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*
Median	28	21	23	19	22	20	23	20	26	21	18	27	15

World Book Company. Booklet containing all tests, 32¢; manual, 16¢; key, 12¢.

Reference: Glenn, E. R. and Rookmyer, I. L. "The Conventional Examination in Chemistry and Physics versus the New Type of Test, Part 3," *School Science and Mathematics*, 23:459–70, May, 1923.

Inventory Test for the Mathematics of High-School Physics

L. R. Kilzer and T. J. Kirby (1929)

Parts I, II

It is somewhat doubtful whether this test belongs here or under mathematics, but this seems the better place. The test is said to serve the three purposes of pointing out to mathematics teachers the items useful for physics, of assisting pupils in deciding whether to take physics or not, and of providing an inventory test to be given a class beginning physics. In Part I are sixty-six problems and questions dealing with arithmetic and algebra, and in Part II twenty-four geometrical exercises, a few of which, however, may be solved by algebraic methods. As a basis for the test the five physics textbooks used most frequently in 345 Iowa high schools were determined and all problems therein solved by every correct method pupils would be

expected to use. A tentative test including all the processes involved except a few that are very rarely employed was tried out and the final test constructed on the basis of the results. Apparently the exercises are arranged in scalar order. Forty minutes are required for each part.

$$r = .90,\ \text{P.E.}_{\text{meas.}} = 3,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .06,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .21.$$

Norms are given as follows:

Percentile	*10*	*25*	*50*	*75*	*90*
Part I	26	30	37	44	50
Part II	9	13	17	20	22
Both	35	44	54	62	71

Public School Publishing Company. Sample set 15¢; $1.00 per 25.

Reference: Kilzer, L. R. "The Mathematics Needed in High School Physics," *Science Education*, 14:335–44, November, 1929.

III. Chemistry

Chemistry along with physics may be considered as one of the somewhat older and better established high-school sciences in contrast with general science and biology. Although not so commonly offered as physics, it has received practically as much attention at the hands of those interested in science tests. One somewhat different tendency is noticeable in their work, however. This is to prepare tests suitable for use in college chemistry classes as well as in high school. Only one such test is described in this section, but there are others designed for college or university use exclusively and, therefore, not described in this chapter.

Harvard High School Chemistry Tests

H. L. Gerry (1922)

Forms A, B

Each test consists of twenty-five exercises, some of which are in multiple-answer, but most in single-answer form. About one

hundred of the items occurring most frequently in the College Entrance Board Examinations for a number of years were tried out and the fifty composing the tests selected. The working time is forty-five minutes, which means that practically all pupils should have as much time as necessary. There are a number of the exercises to which there is not one and only one correct answer, so that, despite the scoring key, scoring is not thoroughly objective. This fault, however, is not great enough to be serious. The exercises are arranged in general order of difficulty, but with several exceptions thereto. One or two are so easy that 90 per cent or more of high-school pupils who have studied chemistry for a year answer them correctly, and the last one so difficult that only about 1 per cent of such pupils are able to answer it. Correlations of these tests with those of Powers, Rauth and Foran, and Rich, range from .33 to .79. The correlation with the composite of the other three is about .75.

$$r = .78,\ \text{P.E.}_{\text{meas.}} = 1.4,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .12,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .32.$$

Medians for about fifteen hundred cases are given as approximately eight for the middle and eleven for the end of the year. Apparently Form A contains items somewhat better learned during the first half year than those in Form B, whereas the reverse is true for the second half year.

Ginn and Company. 80¢ per 30.

References: Gerry, H. L. "Test of High School Chemistry," *Harvard University Bulletin of Education*, No. 9. Cambridge, Massachusetts: Harvard University Press, 1924.

———. "Measurement of the Results of the Teaching of Chemistry," *School Science and Mathematics*, 24:793–804, November, 1924.

———. "The Need and Use of a Scientific Measure of the Results of the Teaching of Chemistry," *School Science and Mathematics*, 25:157–68, February, 1925.

Chemistry Test

S. G. Rich (1923)

Forms Gamma, Epsilon

Each form consists of twenty-five exercises with four possible answers to each. The exercises deal with symbols, equations, re-

actions, problems, general chemical information, and reasons. They are intended to test four phases of the results of instruction, ability to think, information, ability to solve numerical problems, and habits and knowledge acquired from work in the laboratory. Five representative texts, twenty-five recent examinations given by the College Entrance Examination Board and the Regents of the University of the State of New York, and a number of state syllabi were examined for material common to at least two texts or a text, examination and syllabus. Furthermore, items were not chosen unless they appeared to be definitely relevant to at least one of the seven aims of education laid down by the N.E.A. in 1918.[5] The arrangement of exercises is spiral, with cycles of six: thinking, memory, numerical, thinking, memory, laboratory. Of the four suggested answers one is so evidently incorrect that only a person entirely ignorant of chemistry would be likely to choose it, and another nearly enough correct that a person whose knowledge of chemistry was rather vague, or who was careless, might easily select it. Approximately half of the exercises cover material ordinarily taught in the first semester of chemistry, and the other half that taught in the second. The total working time is twenty-five minutes.

On the whole it appears that these tests are fairly satisfactory brief measures of achievement in chemistry. They seem to require more "chemical thinking" and fine distinctions than most others. Also they are rather interesting to pupils. Their reliability is low, however.

$$r = .65,\ \text{P.E.}_{\text{meas.}} = 1,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .08,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .40.$$

Scores correlate about .68 with those on Powers' test, .64 with Gerry's, .65 with Rauth and Foran's, and .67 with the average of the three. The author provides for changing numbers of exercises correct to *T*-scores, and states his norms in terms of the latter. Their approximate equivalents in terms of numbers of exercises correct are also given below, however.

[5] Commission on the Reorganisation of Secondary Education. "Report on Cardinal Principles of Secondary Education," *U. S. Bureau of Education Bulletin*, 1918, No. 35. Washington: Government Printing Office, 1918. 32 p.

Number of semesters studied	*High school*		*College*	
	Number correct	*T-score*	*Number correct*	*T-score*
4	—	—	16.8	60.3
3.5	15.5	56.2	16.0	57.6
3	—	—	16.8	60.3
2.5	14.2	53.3	15.1	55.2
2	14.2	53.3	14.7	54.4
1.5	11.6	48.8	12.6	51.2
1	10.5	44.5	—	—
0.5	10.3	43.6	10.7	45.5

Public School Publishing Company. Sample set 20¢; $1.00 per 25; manual 15¢.

General Chemistry Test

S. R. Powers (1924)

Forms A, B

Part 1 of each form contains thirty multiple-answer items covering general chemical information, and Part 2 thirty-seven exercises dealing with formulae, equations, the chemical names of common substances, and so forth. The arrangement is scalar. Powers tried out 350 items common to most high-school chemistry textbooks and emphasized by most teachers with several thousand pupils as a basis for the test. The working time for the test is thirty-five minutes. Reliability data obtained from experimental forms not identical with the final forms, and only about half as long are:

$$r = .80,\ \text{P.E.}_{\text{meas.}} = 2.5,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .08,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .30.$$

It is likely that the reliability of the published test is higher. Correlations of from .40 to .80 with other chemistry tests and of from .65 to .87 with a composite of three others have been found, also of from .54 to .78 with teachers' marks. On the whole this test is perhaps the best general test in the subject. Norms calculated from scores made on experimental forms by almost three thousand pupils are as follows:

Percentile	*5*	*10*	*25*	*50*	*75*	*90*	*95*
Norm	12	17	26	36	46	55	60

Results reported by Bennett from about half as many pupils in fourteen different cities are on the average three or four points lower at each percentile above the twenty-fifth.

World Book Company. Specimen set 20¢; $1.10 per 25.

References: Powers, S. R. "A Diagnostic Study of the Subject Matter of High School Chemistry," *Teachers College, Columbia University, Contributions to Education*, No. 149. New York: Bureau of Publications, Teachers College, Columbia University, 1924. 84 p.

———— "Achievement in High-School Chemistry: An Examination of Subject Matter," *Teachers College Record*, 15:203–11, May, 1924. Also in *School Science and Mathematics*, 25:53–62, January, 1925.

Bennett, J. C. "A Study of Pupil Errors in Chemistry," *Journal of Educational Research*, 14:275–83, November, 1926.

Symonds, P. M. *Ability Standards for Standardized Achievement Tests in the High School.* New York: Bureau of Publications, Teachers College, Columbia University, 1927. 91 p.

Chemistry Tests

J. W. Rauth and T. G. Foran (1924)

Tests I, II

These two tests deal, respectively, with the work of the first and second semesters of high-school chemistry. They are similar in form, each consisting of four exercises. The first presents the names of twenty substances to be labelled as elements, compounds, or mixtures; the second gives twenty names for which the correct symbols or formulae are to be supplied; the third contains twenty-five true-false statements; and the fourth five numerical problems. The total time required to take Test I is about fifty minutes, and that for Test II slightly longer. The items included were selected after an examination of several widely used textbooks, and agree with the minimum essentials of a high-school chemistry course. The weight given to the fourth exercise is more than one-third that for the whole test, so that five problems count more than half as much as sixty-five other items in determining the total score. This seems to

be giving too high a value to numerical problems in view of the amount of emphasis placed on them by the ordinary teacher.

$$r = .80,\ \text{P.E.}_{\text{meas.}} = 6,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .10,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .30.$$

Scores correlate from .32 to .80 with those on other chemistry tests, and about .78 with a composite of three. For diagnostic purposes these probably rank ahead of a number of the other standarized tests in this subject. Medians for Test I at the end of the first semester and II at the end of the year based on eight hundred or more pupils are as follows:

Test	*Exercise*				
	1	*2*	*3*	*4*	*Total*
I	8	16	13	15	46
II	15	25	20	22	57

The following percentiles are also given:

Test	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
I	27	35	47	59	72	84	90
II	—	—	38	57	74	—	—

Catholic Education Press. $1.00 per 25.

References: Rauth, J. W. and Foran, T. G. "The Rauth-Foran Chemistry Test," *Catholic Education Review*, 22:272–78, May, 1924.

——— "The Rauth-Foran Chemistry Test II," *Catholic Education Review*, 22: 546–50, November, 1924.

Foran, T. G. and Smith, A. V. "Revised Norms for the Rauth-Foran Chemistry Test I," *Catholic Education Review*, 25:286–88, May, 1927.

Smith, A. V. "A Comparative Study of Certain Tests of Achievement in High-School Chemistry," *Catholic University Educational Research Bulletin*, Vol. 2, No. 5. Washington: Catholic Education Press. 45 p.

Columbia Research Bureau Chemistry Test

E. R. Jette, S. R. Powers, and B. D. Wood (1927)

Forms A, B (C to appear later)

Part I on information contains one hundred fifty true-false statements, Part II requires the writing of twenty-two balanced

equations, and Part III calls for the solution of ten problems. The basis of the test is material common to widely used high-school and college texts, and the requirements of college entrance examination agencies. The total working time is an hour and fifty minutes. The reported reliability data are as follows:

$$r = .87,\ \text{P.E.}_{\text{meas.}} = 13,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .09,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .24.$$

The correlation with college chemistry marks is given as .56. Norms for about six thousand Pennsylvania high-school seniors who had just completed one year of chemistry and for a small number who had just completed a year and one-half are as follows:

Years	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
1½	12	38	60	90	138	170	207
1	12	19	38	68	104	142	171

In addition results are given for pupils who have completed one year of chemistry one-half year and a year before taking the test, apparently showing how much is forgotten. Although this ranks high among chemistry tests, this fact is partially due to its unusual length and it is doubtful if the results yielded are any more valid and reliable than those obtained from a combination of two or three other tests which may be given in the same time.

World Book Company. Specimen set 20¢; $1.50 per 25.

Iowa Placement Examinations

Chemistry Aptitude and Training

G. D. Stoddard, Jacob Cornog, C. E. Seashore, and G. M. Ruch

(1924)

Forms A, B

These tests are similar in general form to the others of this series, hence are described along with them on page 380.

Unit Drill Tests in Chemistry
Lyons and Carnahan (1928)
Tests 1–57

This series covers as many units of high-school chemistry as there are tests. Most of the tests contain sixty exercises each, ordinarily in single-answer form. As a basis for selecting the items included, ten textbooks were examined. Each unit dealt with is touched upon in at least three of the ten books, and most of them in all of the ten. It is suggested that on the average about one class period will be required for each complete test, but that ordinarily each pupil will not be asked to respond to all of the exercises. These tests are in no sense standardized and are not intended primarily for testing the results of instruction with a view to determining marks, but rather for use in practice, drill, and diagnostic work. No norms have been established, but it is stated that on the average pupils respond correctly to about 80 per cent of the items. With each test are references to those portions of the ten texts which deal with the unit covered.

Lyons and Carnahan. 1½¢ per copy.

Instructional Tests in Chemistry
E. R. Glenn and L. E. Welton (1929)
Tests 1–36

These tests have passed through four experimental editions tried out during a period of ten years, and should, therefore, be of very high quality. It appears that this expectation is not unfulfilled. Each test covers two pages of a booklet and contains from eighteen up to eighty elements in single-answer, multiple-answer, completion, identification, or some other form. Each deals with a particular topic, of which the following are examples: elements, compounds, and mixtures; measurement of gases; the sodium family; iron, cobalt, and nickel; the technical vocabulary of chemistry. No time limits are provided, but papers are to be collected when all but the three or four slowest in the class have finished. In some cases the time required

should not be more than fifteen or twenty minutes, in others probably a whole class period is necessary.

Since the tests are primarily for instructional rather than measurement purposes, it is suggested that pupils score one another's papers immediately after they have completed the test and then check the scoring of their own. The validity of these tests as determined by a study of their content appears to be decidedly high. Coefficients of reliability of the separate tests range from .70 to .88, the probable errors of measurement from 2 to 5, the ratios of these to the means from .05 to .17, and to the standard deviations from .23 to .36. Percentile norms for fourteen points are given in the manual, but because of the space they would require, are not given below. Median scores are, however, as follows:

Test	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Median	18	30	20	19	16	22	17	16	11	65	23	54

Test	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*
Median	19	26	28	24	21	17	28	16	21	30	19	27

Test	*25*	*26*	*27*	*28*	*29*	*30*	*31*	*32*	*33*	*34*	*35*	*36*
Median	22	24	18	24	27	19	21	27	23	25	20	45

World Book Company. 36¢ per copy; manual 16¢; key 16¢; discount on quantity orders.

IV. Biology

In common with general science the subject of biology is still relatively new in many schools. Formerly it was much more common than now to have a course in botany and perhaps also one in zoölogy, whereas now the tendency is to offer a single unified course under the name of biology. Since this practice has so largely replaced that of a quarter of a century ago, practically all of the available tests in this field are intended for such a course. There is, in so far as the writer knows, just one standard test of botany and none at all of zoölogy.

Biology Test

G. M. Ruch and L. H. Cossmann (1924)

Forms A, B

Although the earliest, this is still one of the best available tests in this subject. In each form are five parts. Subtest 1 contains forty multiple-answer exercises with seven suggested answers to each; Subtest 2 has eighteen similar exercises with three suggested answers to each, many of the answers being complete statements; Subtest 3 requires that the names of fifteen biological structures be matched with the proper portions of drawings; Subtest 4 calls for the filling of four blanks dealing with Mendelian Inheritance; Subtest 5 has five completion paragraphs, with from three to thirteen words omitted in each. The items used in the test were selected as a result of a study of two thousand final examination questions collected from all parts of the country. Three hundred constantly recurring items in the questions were rated by sixty-eight teachers and nine other persons and from the results the final selection was made. The working time is thirty-eight minutes.

Average reliability data are about:

$$r = .82,\ \text{P.E.}_{\text{meas.}} = 3.5,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .10,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .28.$$

Norms for high-school pupils, presumably mostly sophomores, are given below. Those for the end of the first semester are probably based on too few pupils to be very reliable, but those for the end of the year are based on a larger number:

	Percentile				
	10	*25*	*50*	*75*	*90*
Year	23	30	40	53	64
Semester	10	17	25	34	48

World Book Company. Specimen set 20¢; $1.30 per 25.

References: Ruch, G. M. and Cossmann, L. H. "Standardized Content in High School Biology," *Journal of Educational Psychology*, 15:285–96, May, 1924.

Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction.* Yonkers, New York: World Book Company, 1927, p. 142–44.
Symonds, P. M. *Ability Standards for Standardized Achievement Tests in the High School.* New York: Bureau of Publications, Teachers College, Columbia University, 1927. 91 p.

Information Exercises in Biology

J. L. Cooprider (1924)

The exercises are six in number, each in a different form. The first contains seventeen completion sentences; the second, sixteen true-false statements; the third, seventeen multiple-answer items; the fourth, nine best-reason statements; the fifth, eighteen classification exercises; and the last, seventeen demanding logical selection. As the name of this test implies, the exercises are informational and appear to sample first-year biology fairly well. They are based upon two surveys of textbooks and much experimental work in trying out exercises. About three-eighths of them pertain to zoölogy, one-fourth to botany, one-fifth to human biology and the remainder are general. No time limit is given, but it is suggested that pupils should not use more than forty minutes, and that most need only twenty-five or thirty. The list of correct answers is not in the most convenient form for use. The coefficient of reliability is .92 and the correlation with teachers' estimates .70. Mean scores based on about six hundred pupils are given as follows:

Pupils who have not studied biology, 29
After one semester of biology, 46
After one year of biology, 58

Public School Publishing Company. Sample set, 10¢; 50¢ per 25.

Catholic High School Test in Biology

L. B. Jordan and T. G. Foran (1926)

Forms A, B

The first part of each form contains forty multiple-answer statements; the second, eighteen multiple-reason statements; the third requires the matching of nineteen drawings with the things

represented; the fourth is in completion form, consisting of about thirty words to be supplied. The total working time is forty-five minutes, apparently enough that speed is not an important factor in determining the score. Median scores at the end of the year for the four parts are 26, 14, 25 and 17, respectively. For total scores the first quartile is 64, the median 78 and the third quartile 89.

Catholic Education Press. $1.00 per 25.

Virginia Biology Test

F. S. King (1928)

Forms A, B

This test consists of five parts, in multiple-choice, true-false, matching, best-answer, and completion form, respectively. The total number of pupil responses called for is almost two hundred. It was based upon the collection of examination papers made by Ruch and Cossmann, and utilized for their test also. King further checked each item with Peabody and Hunt's *Biology and Human Welfare*, and with the Virginia State Biology Course of Study. From the items found common, fifteen hundred questions, were developed, criticized, and the ones that appeared to be most suitable tried out. Finally, on the basis of the results obtained they were revised, arranged in order of difficulty in Subtests 1, 2, and 4, and the present form issued. This procedure would appear to result in high curricular validity for Virginia, at least, and probably fairly high for other states. The time is forty-five minutes. For a few pupils this test correlates .68 with Cooprider's. Average reliability data are:

$$r = .83,\ \text{P.E.}_{\text{meas.}} = 4,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .04,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .28.$$

The fifth subtest, the one in completion form, is not perfectly objective, but the subjective element is not great enough to be a serious defect. The most recent norms published are based on well over a thousand pupils from more than fifty Virginia high schools for the end of the year.

Percentile	*5*	*10*	*25*	*50*	*75*	*90*	*95*
Score	65	73	91	113	136	153	163

For a few hundred pupils norms for the end of the first semester are: first quartile 70, median 87, third quartile 107.

Bureau of Tests and Measurements. Specimen set 20¢; 4¢ per copy.

Reference: King, F. S. "An Account of the Derivation and Standardization of the Virginia Biology Test," *Virginia Journal of Education*, 22: 161–66, December, 1928.

Test of General Biology

M. E. Oakes and S. R. Powers (1929)

Forms A, B

The three sections test knowledge of a total of one hundred terms, most of which are embodied in multiple-answer exercises but some in alternative form. The selection of material included was based on a study of New York State Regents Examinations for almost twenty years, two studies of the content of high-school biology courses and texts, and several others having to do with what the content of a biology course should be. The almost six hundred items resulting from this study were tried out and reduced to two hundred. Those in each test range from one item so difficult that only 4 per cent of the pupils upon whom the material was tried out answered it correctly, up to one or more so easy that more than 90 per cent responded correctly. The items do not appear to be in order of difficulty, however. It is not intended that rigid time limits be employed, but thirty-five minutes are stated to be adequate. The function of the test evidently is to yield a general measure of achievement in high-school biology rather than to be to any considerable extent diagnostic. The care used in the choice of items was such that its validity for this purpose seems decidedly high. Coefficients of reliability averaging slightly above .90 are reported for the preliminary forms. It is stated that a score of fifty is the norm. Apparently this is for the end of the year.

Bureau of Publications. Specimen set, 10¢; $2.00 per 100.

Biology Tests

T. R. Stemen and W. S. Myers (1930)

Tests 1, 2, 3, 4, 5, 6

Each of these tests includes about one hundred elements grouped in several parts and covering the work of one six-weeks' period. Multiple-answer, true-false, matching, and completion exercises are employed. Forty minutes are allowed for taking the test.

Harlow Publishing Company. Sample set 10¢; 75¢ per 25, $2.50 per 100.

Instructional Tests in Biology

J. G. Blaisdell (1929)

Tests 1–25

This series of tests contains more than twelve hundred elements dealing with animal, human, and plant biology, and so chosen as to be appropriate for use with any of the most widely used texts and also state and city courses of study. Each deals with a single topic and is intended to be given when that has been completed. Three of the twenty-five are summary tests covering the general topics of animal, human, and plant biology, and one a general test of all three fields. It is suggested that these summary tests may be employed as final examinations. The exercises employed include multiple-answer, completion, identification, true-false, matching, and other types. For three of the tests the time allowance is thirty minutes each, for four it is ten minutes, and for each of the others twenty. Percentile norms based on about five hundred cases from several states are given for each of the tests. Those for the summary tests may be found below:

Test	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
Animal biology	45	51	56	62	67	72	74
Human biology	36	43	50	58	67	71	74
Plant biology	27	33	44	55	63	69	74
General biology	15	20	29	35	41	45	48

World Book Company. 32¢ per set; manual, 12¢.

Tests to Accompany a Biology Workbook

J. C. Adell, Orra O. Durham, and L. E. Welton 1929)

Units I–XVI; Tests 1, 2 on some, 1 only on others

As their name implies, this series of tests is intended to accompany a biology workbook, and especially the one by the same authors. The tests are fitted to the sixteen units of this book, there being either one or two tests on each. Each test is a single sheet calling for from six to thirty-three responses. Some are in picture or identification form, others true-false and completion. Apparently there are no time limits.

Ginn and Company. $1.80 per 10 of complete series.

Michigan Botany Test

O. W. Laidlaw and Clifford Woody (1925)

This, the only standardized test in the subject, is a revision of an earlier test by the same authors. It contains four subtests. The first three call for twenty responses each, and are in true-false, multiple-answer and matching form, respectively. The fourth contains ten multiple-reason exercises. The time is twenty-five minutes.

$$r = .87,\ \text{P.E.}_{\text{meas.}} = 3,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .07,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .25,$$

The correlation with term marks is .53. For the end of the year $Q_1 = 35$, Md. $= 45$, $Q_3 = 53$, and for the end of one semester Md. $= 37$.

Public School Publishing Company. Sample set 15¢; $1.00 per 25.

V. Miscellaneous

In addition to the tests in the specific subjects classed under high-school science there have also been at least two others which seem to deserve brief mention. The first of these, by Herring,[6] is not and never has been commercially available, but

[6] Herring, J. P. "Measurements of Some Abilities in Scientific Thinking," *Journal of Educational Psychology*, 9:535–58, December, 1918.

———. "Derivation of a Scale to Measure Abilities in Scientific Thinking," *Journal of Educational Psychology*, 10:417–32, November, 1919.

seems worthy of mention because of its suggestiveness. It consisted of thirty-three exercises three of which deal with each of the following phases of scientific thinking: classification, clarity, relevancy, definition, feasibility, arrangement, recording, sufficiency, value, comparison, statistics. The form of exercise employed is the multiple-answer. Data on reliability, validity, and so forth, were worked out but will not be given here.

The second test of the two referred to is actually commercially available, although it will probably be used in college rather than in high school. It is described below.

Stanford Scientific Aptitude Test
D. L. Zyve (1929)

This test consists of eleven exercises following a preliminary questionnaire. Most of these are in multiple-answer form, although some are problems and one or two something else. They deal to some extent with scientific information that might likely be picked up outside of regular school courses, with judgment and reasoning, clarity of definition, accuracy of observation and interpretation, and so forth. Two hours are allowed for the test, which should be enough to eliminate the speed element for practically all those taking it. Correlations of scores on this with intelligence-test scores range from .50 down. Those with school marks tend to be about the same. Ratings on scientific aptitude given by a group of judges, however, correlate .74 with test scores. The validity of the test is stated to be represented by a coefficient of .82. The figures for reliability seem not to be entirely consistent. Apparently they are as follows:

$$r = .74,\ \text{P.E.}_{\text{meas.}} = 8,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .06,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .34.$$

Scores for small groups of college students and faculty groups are as follows:

Unselected freshmen	105
Science freshmen	113
Non-science seniors and graduates	90
Science and engineering students	134
Non-science faculty	184
Science faculty	153

Stanford University Press. 10¢ per copy, $2.00 per 10, $3.50 per 20, $7.00 per 50, $14.50 per 100, $70.00 per 500; explanatory booklet 25¢; directions and key 25¢.

BIBLIOGRAPHY

I. General Science

Caldwell, O. W., et al. "Reorganization of Science in Secondary Schools," *U. S. Bureau of Education Bulletin*, 1920, No. 26. Washington: Government Printing Office, 1920. 62 p.

Cunningham, H. A. "Character and Value of Existing Tests, for Pupils and Teachers, in General Science," *General Science Quarterly*, 13:61–69, January, 1929.

Curtis, F. D. "Some Values Derived from Extensive Reading of General Science," *Teachers College, Columbia University, Contributions to Education*, No. 163. New York: Bureau of Publications, Teachers College, Columbia University, 1924. 142 p.

Downing, E. R. "A Range of Information Test in Science," *School Science and Mathematics*, 19:228–33, March, 1919.

———. "The Range of Information Test in Science Revised," *School Science and Mathematics*, 20:77–83, January, 1920.

———. "The Revised Norms for the Range of Information Test in Science," *School Science and Mathematics*, 26:142–46, February, 1926.

Eckert, D. Z. "Report on Tests in History and General Science at Close of First Semester," *Curriculum Study and Educational Research Bulletin (Pittsburgh Public Schools)*, 3:178–86, March–April, 1929.

Glenn, E. R. and Walker, Josephine. "A Bibliography of Science Teaching in Secondary Schools," *U. S. Bureau of Education Bulletin*, 1925, No. 13. Washington: Government Printing Office, 1925. 161 p.

Maxwell, P. A. "Tests in General Science," *General Science Quarterly*, 4: 443–50, May, 1920.

Odell, C. W. "Scales for Rating Pupils' Answers to Nine Types of Thought Questions in General Science," *General Science Quarterly*, 12:317–28, 382–90, 467–76, 524–36; November, 1927, January, March, May, 1928.

———. "The Use of Scales for Rating Pupils' Answers to Thought Questions," *University of Illinois Bulletin*, Vol. 26, No. 36, Bureau of Educational Research Bulletin No. 46. Urbana: University of Illinois, 1929. 34 p.

Ottmyer, E. F. "Results of an Objective Standard Test on Weather," *General Science Quarterly*, 8:500–4, March, 1924.

Powers, S. R. "The Vocabularies of High School Science Textbooks," *Teachers College Record*, 26:368–83, May, 1925.

Rich, S. G. "The Available Tests for Results of Teaching the Sciences," *School Science and Mathematics*, 26:845–52, November, 1926.

Ruch, G. M. "Tests and Measurements in High School Science," *School Science and Mathematics*, 23:885–91, December, 1923.

Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction.* Yonkers, New York: World Book Company, 1927, p. 136–41.
Symonds, P. M. *Measurement in Secondary Education.* New York: The Macmillan Company, 1927, p. 130–33.
Toops, H. A. "A General Science Test," *School Science and Mathematics,* 25:817–22, November, 1925.
Watkins, R. K. "The Technique and Value of Project Teaching in General Science," *General Science Quarterly,* 7:235–56, May, 1923; 8:311–41, 387–422, November, 1923, January, 1924.

II. Physics

Bear, R. M. "The Predictive Value of the Iowa Physics Aptitude Placement Test," *Journal of Applied Psychology,* 11:381–84, October, 1927.
Burgess, T. O. "A Psychological Analysis of Abilities in High School Physics," *University of Iowa Studies in Education,* Vol. 3, No. 6. Iowa City: University of Iowa, 1926. 24 p.
Camp, H. L. "An Evaluation of Standard Tests and Suggested Uses in Improving Physics Teaching," *School Science and Mathematics,* 23:441–46, May, 1923.
Chapman, J. C. "The Measurement of Physics Information," *School Review,* 27:748–56, December, 1919.
Franzén, C. G. F. "An Experiment in the Content of High School Physics," *Fourteenth Indiana University Educational Measurements Conference.* Bloomington: Bureau of Coöperative Research, Indiana University, 1927, p. 42–45.
Glenn, E. R. "The Conventional Examination in Chemistry and Physics versus the New Types of Tests," *School Science and Mathematics,* 21: 666–70, 746–56, October, November, 1921.
Glenn, E. R. and Heck, A. O. "Preliminary Studies of Achievement in Large City High Schools," *Contributions to Education,* Vol. I. Yonkers, New York: World Book Company, 1924, Chapter XXX.
Hawthorne, W. C. "Standardized Tests in Physics," *High School Conference Proceedings,* 1923. Urbana: University of Illinois, 1924, p. 352–57.
Hurd, A. W. "Reorganization in Physics," *North Central Association Quarterly,* 4:277–93, September, 1929.
Jones, F. T. "Practice Exercises in Physics," *School Review,* 26:341–48, May, 1918.
Peters, C. J. and Watkins, R. K. "Objective Tests for High School Physics." Columbia, Missouri: C. J. Peters and R. K. Watkins, 1926. 36 p.
Randall, D. P., Chapman, J. C., and Sutton, C. W. "The Place of the Numerical Problem in High-School Physics," *School Review,* 26:39–43, January, 1918.
Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction.* Yonkers, New York: World Book Company, 1927, p. 152–56.
Smith, Gale. *Twentieth Century Practice Exercises and Objective Tests in Physics.* Fowler, Indiana: Benton Review Shop, 1929. 113 p.

Starch, Daniel. *Educational Measurements.* New York: The Macmillan Company, 1916, p. 188–93.

Symonds, P. M. *Measurement in Secondary Education.* New York: The Macmillan Company, 1927, p. 135–38.

Thorndike, E. L. "Completion Tests in Physics," *School Science and Mathematics,* 22:637–47, October, 1922.

"Objective Achievement Tests Constructed and Used in St. Louis," (*St. Louis*) *Public School Messenger,* 25:107–13, November 30, 1927.

III. Chemistry

Bell, J. C. "A Study of Attainments of High-School Pupils in First-Year Chemistry," *School Science and Mathematics,* 18:425–32, May, 1918.

————. "A Test in First-Year Chemistry," *Journal of Educational Psychology,* 9:199–209, April, 1918.

Cook, C. G. *New Type Questions in Chemistry.* New York: Globe Book Company, 1927. 91 p.

Cornog, Jacob and Stoddard, G. D. "Predicting Performance in Chemistry," *Journal of Chemical Education,* 2:701–9, August, 1925.

Garner, Edith. "A Study of Chemistry Examination Questions Given by Various States and Cities in the Middle West and East," *School Science and Mathematics,* 27:140–43, February, 1927.

Gerry, H. L. "College Entrance Examination Board Questions in Chemistry," *School Science and Mathematics,* 20:845–50, December, 1920.

————"Types of Tests Desirable for Chemistry and the Present Status of Their Development," *School Science and Mathematics,* 25:918–22, December, 1925.

Glenn, E. R. "Conventional Examinations in Chemistry and Physics versus the New Types of Tests," *School Science and Mathematics,* 21:666–70, 746–56, October, November, 1921.

Glenn, E. R. and Welton, L. E. *New Types of High School Chemistry Tests for Instructional Purposes.* New York: E. R. Glenn, 1922. 76 p.

Guilford, J. P. and Hyde, W. F. "A Test for Classification of Students in Chemistry," *Journal of Applied Psychology,* 9:196–202, June, 1925.

Hawthorne, W. C. "Standardized Tests," *School Science and Mathematics,* 23:791–98, November, 1923.

Hayes, Seth. "Coöperative Chemistry Tests," *Journal of Educational Research,* 9:109–20, September, 1921.

Mabee, F. C. "A Test of Achievement in College Chemistry and Results Obtained from Its Use with Both High-School and College Classes," *Journal of Chemical Education,* 3:70–76, January, 1926.

Powers, S. R. "The Achievement of High-School Students and College Freshmen in Chemistry," *School Science and Mathematics,* 21:366–77, April, 1921.

————. "Tests of Achievement in Chemistry," *Journal of Chemical Education,* 1:139–44, September, 1924.

————. "A Comparison of the Achievement of High School and Uni-

versity Students in Certain Tasks in Chemistry," *Journal of Educational Research*, 6:332–43, November, 1922.

————. "How Long Do Students Retain What They Have Learned from High School Chemistry?" *Contributions to Education*, Vol. 1. Yonkers, New York: World Book Company, 1924, Chapter XXXI.

Rich, S. G. "Achievements of Pupils in Chemistry," *School Science and Mathematics*, 25:145–49, February, 1925.

Rivett, B. J. "A Comprehensive Chemistry Test," *School Science and Mathematics*, 23:377–86, April, 1923.

————. "Results with Standard Chemistry Tests," *School Science and Mathematics*, 21:720–22, November, 1921.

————. "Testing Results in Chemistry," *School Science and Mathematics*, 19:742–45, November, 1919.

Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction*. Yonkers, New York: World Book Company, 1927, p. 146–52.

Smith, A. V. "A Comparative Study of Certain Tests of Achievement in High School Chemistry," *Catholic University, Educational Research Bulletins*, Vol. 2, No. 5. Washington, D. C.: Catholic Education Press, 1927. 45 p.

Stout, L. E. "The Selective Value of Powers' General Chemistry Test, Scale 'A,'" *Journal of Chemical Education*, 3:1138–43, October, 1926.

Symonds, P. M. *Measurement in Secondary Education*. New York: The Macmillan Company, 1927, p. 139–45.

Webb, H. A. "A Preliminary Test in Chemistry," *Journal of Educational Psychology*, 10:36–43, January, 1919.

"Objective Achievement Tests Constructed and Used in St. Louis," (*St. Louis*) *Public School Messenger*, 25:133–48, November, 1927.

IV. Biology

Grier, N. M. "The Range of Information Test in Biology," *Journal of Educational Psychology*, 9:210–16, 388–92, April, September, 1918; 10:509–16, December, 1919.

Hunter, O. B. and Moss, F. A. "Standardized Tests in Bacteriology," *Public Personnel Studies*, 3:52–66, February, 1925.

Richards, O. W. "Test Construction in Less Standardized Subjects Illustrated by the Richards Biology Test," *School Science and Mathematics*, 27:22–27, January, 1927.

Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction*. Yonkers, New York: World Book Company, 1927, p. 141–46.

Symonds, P. M. *Measurement in Secondary Education*. New York: The Macmillan Company, 1927, p. 133–35.

CHAPTER IX

SOCIAL STUDIES

Introduction.—Although the total number of tests produced in history and the other social studies is large, ranking next to arithmetic, reading, and language and grammar, yet the number of such tests appropriate for high-school use is not so great as might be expected. Needless to say, much of the activity in this field has been at the elementary-school level. Probably the most important cause of the relatively small number of tests suitable for high-school use, however, is that many of the efforts at test construction in these subjects have not been carried to completion, and that a larger proportion of the tests actually for sale seem to possess relatively little merit than is true in many other subjects.

In the social studies probably more than in any of the more common subjects except literature, teaching emphasis is or should be placed on other outcomes than mere information or knowledge. Historical interpretation and judgment, the ability to apply the lessons of the past to the present and future, the proper attitudes and ideals with regard to citizenship, and other similar outcomes, occupy an important place in most lists of objectives, even though they may not in actual teaching practice. Such results are, however, difficult to measure, so that most of the actually available tests are little if any more than tests of information acquired. A number of them have attempted to measure other phases, but in general have not been very successful in so doing. For example, several tests have given attention to the measurement of character judgment, usually by requiring that the proper adjectives to describe the characters named be selected. In many cases pupils respond to such exercises by recalling the adjectives applied to the given characters in the text, and thus exercise no judgment whatsoever, but merely memory. The same is true with regard to most attempts to meas-

ure other outcomes than merely information. It seems that if these abilities are to be measured satisfactorily the exercise should deal with suppositious characters and situations so that it will be impossible for pupils to respond correctly by the exercise of mere memory. Unfortunately practically no tests based upon such content have been devised.

It might be inferred from what has just been said that the factual side of history is not important. This, however, is not true. Before an individual can think in historical terms he must possess the proper data upon which to base his thinking. Moreover for purposes of general culture, understanding of allusions in literature, and so forth, it is important that one be in possession of a fair stock of historical facts. There is, therefore, a place for tests that deal with knowledge of dates, characters, events, and so forth, provided it is realized in using them that what they measure is not the sole nor chief desirable outcome of the study of history. Furthermore, on the whole there is a positive correlation between historical knowledge and these other more intangible outcomes so that tests of the sort actually available are to some extent indicative of these other phases.

One criticism that applies to many of the available tests is that they cover the whole year's work and thus are not entirely suitable for use until the end of the year. If such tests are given during the year when pupils have studied only a portion of the content which they cover, scores made will depend to a considerable degree not only upon how well pupils have mastered the assigned work, but also upon how much information concerning the portion of the work not yet covered they have picked up from general reading and other sources. Thus the test scores are not valid measures of what has been learned in class but rather of total historical knowledge regardless of where obtained. As will be seen from the test descriptions, an effort has been made to meet this criticism in a few cases by constructing series of two or more tests, each covering a portion of the whole course.

I. American History

Since American history is required in a large majority of high schools and is offered in practically all others, it has received

CHAPTER IX

SOCIAL STUDIES

Introduction.—Although the total number of tests produced in history and the other social studies is large, ranking next to arithmetic, reading, and language and grammar, yet the number of such tests appropriate for high-school use is not so great as might be expected. Needless to say, much of the activity in this field has been at the elementary-school level. Probably the most important cause of the relatively small number of tests suitable for high-school use, however, is that many of the efforts at test construction in these subjects have not been carried to completion, and that a larger proportion of the tests actually for sale seem to possess relatively little merit than is true in many other subjects.

In the social studies probably more than in any of the more common subjects except literature, teaching emphasis is or should be placed on other outcomes than mere information or knowledge. Historical interpretation and judgment, the ability to apply the lessons of the past to the present and future, the proper attitudes and ideals with regard to citizenship, and other similar outcomes, occupy an important place in most lists of objectives, even though they may not in actual teaching practice. Such results are, however, difficult to measure, so that most of the actually available tests are little if any more than tests of information acquired. A number of them have attempted to measure other phases, but in general have not been very successful in so doing. For example, several tests have given attention to the measurement of character judgment, usually by requiring that the proper adjectives to describe the characters named be selected. In many cases pupils respond to such exercises by recalling the adjectives applied to the given characters in the text, and thus exercise no judgment whatsoever, but merely memory. The same is true with regard to most attempts to meas-

ure other outcomes than merely information. It seems that if these abilities are to be measured satisfactorily the exercise should deal with suppositious characters and situations so that it will be impossible for pupils to respond correctly by the exercise of mere memory. Unfortunately practically no tests based upon such content have been devised.

It might be inferred from what has just been said that the factual side of history is not important. This, however, is not true. Before an individual can think in historical terms he must possess the proper data upon which to base his thinking. Moreover for purposes of general culture, understanding of allusions in literature, and so forth, it is important that one be in possession of a fair stock of historical facts. There is, therefore, a place for tests that deal with knowledge of dates, characters, events, and so forth, provided it is realized in using them that what they measure is not the sole nor chief desirable outcome of the study of history. Furthermore, on the whole there is a positive correlation between historical knowledge and these other more intangible outcomes so that tests of the sort actually available are to some extent indicative of these other phases.

One criticism that applies to many of the available tests is that they cover the whole year's work and thus are not entirely suitable for use until the end of the year. If such tests are given during the year when pupils have studied only a portion of the content which they cover, scores made will depend to a considerable degree not only upon how well pupils have mastered the assigned work, but also upon how much information concerning the portion of the work not yet covered they have picked up from general reading and other sources. Thus the test scores are not valid measures of what has been learned in class but rather of total historical knowledge regardless of where obtained. As will be seen from the test descriptions, an effort has been made to meet this criticism in a few cases by constructing series of two or more tests, each covering a portion of the whole course.

I. American History

Since American history is required in a large majority of high schools and is offered in practically all others, it has received

far more attention at the hands of test makers than have other history courses. At least six times as many tests have appeared in it as in European and general history combined. Many of the tests are intended for use in both the seventh and eighth grades and high school, ordinarily the junior and senior years, since American history is commonly taught at those times.

Diagnostic Tests in American History
A. S. Barr (1920)
Series A, B

Each series consists of a full-page practice test followed by five subtests dealing with comprehension, chronological judgment, historical evidence, evaluation of facts, and causal relationships. Within each subtest are from three to nine exercises, some of which are subdivided, calling for responses in multiple-answer form. The tests, as their name implies, are intended to be diagnostic, and are so with regard to general historical ability or understanding rather than information actually acquired. The items included are based upon Bagley and Rugg's study of twenty-three texts used in American history,[1] and probably because of this fact neglect recent years. The test passed through several revisions before its present form appeared, but these did not suffice to relieve it of a rather complicated system of weighting scores. Fractions with eight different denominators need to be combined somewhere in the scoring process. The time required is about a class period. These tests yield low correlations with other history tests, also with school marks. Average figures from several studies of reliability are about:

$$r = .74,\ \text{P.E.}_{\text{meas.}} = 3.5,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .07,\ \frac{\text{P.E.}_{\text{meas.}}}{\quad} = .35.$$

Senior high-school norms are given as follows:

[1] Bagley, W. C. and Rugg, H. O. "The Content of American History as Taught in the Seventh and Eighth Grades," *University of Illinois Bulletin*, Vol. 13, No. 52, School of Education Bulletin No. 16. Urbana: University of Illinois, 1916. 59 p.

Form	Test				
	1	2	3	4	5
2A	8.2	7.5	6.2	12.2	8.9
2B	10.0	8.0	9.0	12.3	7.5

Public School Publishing Company. Sample set 20¢; 5¢ per copy, $4.00 per 100.

References: Odell, C. W. "The Barr Diagnostic Tests in American History," *School and Society*, 16:501–3, October 28, 1922.
Wessel, H. M. "A Critical Study of the Results of the Barr's Diagnostic Tests in American History," *Historical Outlook*, 17:230–31, May, 1926.

American History Scales
M. J. Van Wagenen (1922)
Information Scale S–3

The first series of history scales constructed by Van Wagenen consisted of scales dealing with information, thought, and character judgment, but these were later revised and a series of eleven information scales and one thought scale published. Of these only one is intended for high-school use. It is a general scale containing thirty exercises, almost all multiple-answer, dealing with American history from the earliest times to the present. The exercises are arranged in order of difficulty at intervals of one from seventy-one up to one hundred, inclusive. These weights were very carefully determined by trying out the exercises. Apparently the test yields a rather good general measure of knowledge of American history, but is too short and general to have diagnostic value. Scoring is somewhat complicated, a score of a given amount denoting the ability to do half of those at the given difficulty correctly. The test can easily be given within a forty-minute period.

$$r = .76,\ \text{P.E.}_{\text{meas.}} = 3,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .04,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .33.$$

An average correlation of .60 with five other history tests is reported. A tentative norm of ninety-two for high-school freshmen is suggested.

Bureau of Publications. Specimen set 50¢; $2.00 per 100; manual 35¢.

References: Van Wagenen, M. J. "Historical Information and Judgment in Pupils of Elementary Schools," *Teachers College, Columbia University, Contributions to Education*, No. 101. New York: Bureau of Publications, Teachers College, Columbia University, 1919. 74 p.

————."Some Implications of the Revised Van Wagenen History Scales," *Teachers College Record*, 27: 142–48, October, 1925.

Symonds, P. M. *Ability Standards for Standardized Achievement Tests in the High School*. New York: Bureau of Publications, Teachers College, Columbia University, 1927. 91 p.

Test on Understanding of American History

L. W. Pressey and R. C. Richards

Although this rather simple test is apparently better suited for use in elementary than in high school, it has received some use in the latter, and appears to merit mention here. It consists of four subtests dealing with character judgment, historical vocabulary, sequence of events, and cause and effect relationships. Each calls for twenty-five responses to multiple-answer exercises. The elements included are those found common to six widely used American history textbooks. A number of teachers of history participated in selecting the adjectives used in the character-judgment test. If the character-judgment test is omitted, since the items therein are rather difficult to classify, approximately one-third of the remaining elements are devoted to political matters, the same proportion to social and economic matters, and about one-fifth to military events. The total working time allowed is twenty-five minutes, which should be sufficient that speed is not an important factor.

$$r = .89,\ \text{P.E.}_{\text{meas.}} = 3,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .05,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .22.$$

Its average correlation with five other history tests is .67. For October testing the following norms are announced for high-school seniors:

Subtest	*1*	*2*	*3*	*4*	*Total*
Norm	15	17	16	15	63

This is one of the simplest, most easily administered, understood and scored of the history tests available. It is very doubtful, however, if it is really as diagnostic as is implied by the subtest titles and if the authors' interpretations of scores are justified.

Public School Publishing Company. Sample set 10¢; $2.00 per 100.

Reference: Pressey, L. C. "Standard Tests in the Understanding of American History," *Educational Research Bulletin (Ohio State University)*, Vol. 3, No. 2. Columbus: Bureau of Educational Research, Ohio State University, January 23, 1924, p. 28–31.

Tests in American History

C. A. Gregory (1923)

Test III; Forms A, B

Although a revised series has appeared, this one of Gregory's original tests has not been changed. Each form is composed of seven parts as follows:

1. Miscellaneous facts and dates
2. The period of discovery, exploration, and colonization
3. The period of the Revolution, from 1760 to 1789
4. The period of national growth, from 1789 to 1830
5. The period of sectional disputes and Civil War, 1829 to 1865
6. The period of reconstruction and national development, from 1865 to 1900
7. The period from 1900 to 1922

The first part includes forty completion statements and each of the others ten multiple-answer exercises. It is evident that this number is too small to yield more than a very general measure of pupils' achievement in each of the phases of history dealt with. It is recommended, therefore, that this test only be used when a general measure is desired. For this purpose, however, it ranks high. Pupils are to be given all the time needed to do the test, but it should not require more than an ordinary high-school period, and indeed thirty minutes will probably be sufficient

for most pupils. Reliability data differ, but average about as follows:

$$r = .85,\ \text{P.E.}_{\text{meas.}} = 3,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .07,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .26.$$

The average correlation with five other history tests is .65. The norm for Grade VIII is 43, for XI 48 and for XII 57.

Bureau of Administrative Research. Specimen set 10¢; 4¢ per copy, $3.50 per 100.

American History Exercises

J. F. Tyrrell (1925)

Tests 1–48

This series of non-standardized tests covers the whole period of American history. Each contains fifty elements. They are grouped in sets of three, there being one true-false, one completion, and one recognition or multiple-answer test on each of sixteen periods, the first being that of exploration and the last covering the past thirty years or so. The tests appear to have some value for practice and study purposes, since they include items of information commonly found in textbooks and taught by teachers. Regarded as a whole they emphasize unduly earlier portions of our history as contrasted with that following the Civil War, since of the sixteen periods dealt with only two fall within the later time. About fifteen minutes are needed for each.

Palmer Company. Sample set 75¢; 25¢ per 25; $1.35 per 50 of each of three tests on one period.

Columbia Research Bureau American History Test

H. J. Carman, T. N. Barrows, and R. D. Wood (1926)

Form A, B

Each form has four subtests. In the first are eighty true-false statements, in the second eight matching exercises, in the third fifty multiple-answer exercises, and in the fourth twenty completion exercises. The total working time allowed is ninety minutes. The elements included were carefully apportioned

among the different periods of American history, half, for example, belonging to the time before the Civil War, and half to that since, and also according to five phases of history. On the latter basis about half deal primarily with political material, a fourth with economic, a seventh with social, and a tenth with religious and educational. For the preliminary forms

$$r = .91,\ \text{P.E.}_{\text{meas.}} = 4.5,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .04,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .20.$$

For the final forms reliability is probably higher. In one large school test scores correlated .82 with final marks and .47 with Regents examination grades. Norms based on one group of over six hundred high-school seniors at the end of American history and less than two hundred entering college freshmen who had credit in American history, and another of almost thirty-seven hundred Pennsylvania high-school seniors are:

	Percentile						
	5	*10*	*25*	*50*	*75*	*90*	*95*
Seniors and freshmen	68	81	96	111	127	140	147
Pennsylvania seniors	33	44	60	68	75	121	138

World Book Company. Specimen set 30¢; $1.50 per 25.

Iowa General Information Test in American History
M. H. DeGraff, G. M. Ruch, and H. A. Greene (1927)
Forms A, B

Each form contains a total of one hundred completion statements. These tests are the outgrowth of the Commonwealth Investigation of Objective Examinations in the Social Studies conducted by Ruch and others.[3] Their content was selected from three widely used and generally accepted textbooks, Muzzey's *American History*, Gordy's *History of the United States*, and Beard and Bagley's *History of the American People*. The items included were more or less common to these three and formed the basis for considerable discussion therein. Three hundred fifty

[3] Ruch, G. M., et al. *Objective Examination Methods in the Social Studies.* Chicago: Scott, Foresman and Company 1926. 116 p.

such items which could be stated concisely and without ambiguity and to which objective answers were possible were formulated, and as a result of trying them out two lists of a hundred each were chosen.

For high-school pupils it is stated that forty-one minutes are required to allow all pupils to finish, whereas thirty-three or thirty-four minutes allow 90 per cent to finish, thirty or thirty-one, 80 per cent, and so on to eighteen or twenty minutes for 10 per cent. For tests of only one hundred items, and indeed by comparison with most standardized tests, their reliability is high.

$$r = .95,\ \text{P.E.}_{\text{meas.}} = 2.5,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .09,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .15.$$

Mean scores for the end of the year for high-school pupils completing American history are thirty-eight for Form A and two or three points higher for Form B. Other data, however, indicate that the two forms are more nearly equivalent than these norms indicate.

The tests possess one decidedly unusual feature. Answers are not written in the test booklets themselves, but in specially prepared narrow strips which are laid alongside the exercises. Thus the test booklets can be used over and over again by securing additional strips at very slight expense.

Bureau of Educational Research and Service. 15¢ per copy; manual 15¢; answer strips $1.00 per 100.

Study Guide Tests in American History

M. J. Stormzand (1927)

Parts I, II

These tests, intended for study or drill rather than measurement purposes, have been prepared for use in connection with Beard and Bagley's *History of the American People*. Each part is published in a booklet with perforated leaves so that separate tests can be removed if desired. The first deals with the period from 1492 to 1860, and the second with that from 1815 to the present. There are thirty-two sheets, practically all printed on both sides, in each part. Of these the last nine in the first part and

the first nine in the second, are identical. The test exercises are in various forms, completion, enumeration, yes-no, single-answer, map location, matching, and so forth, all being such that answers are relatively, if not absolutely, objective. On the average each sheet will probably require twenty to thirty minutes to answer the exercises contained thereon.

The Macmillan Company. Each part, 36¢ per copy.

Reference: Stormzand, M. J. "American History Teaching and Testing. Supervised Study and Scientific Testing in American History, Based on Beard and Bagley's *The History of the American People*." New York: The Macmillan Company, 1926. 181 p.

Twenty Tests to Accompany Muzzey's *History of the American People*

Mildred C. Bishop and E. K. Robinson (1929)

Tests I–XX

This is a series of non-standardized tests each of which covers one section, usually about thirty or forty pages, of the text mentioned above. They include completion, true-false, multiple-answer, and matching exercises. The number of responses called for in the different tests varies, but is usually about twenty. Apparently ten minutes should be enough for any one of the series.

Ginn and Company. 20¢ per set.

Students' Objective-Test Manual to Accompany Muzzey's *History of the American People*

H. C. Perkins (1930)

Assignments 1–140; Topical Tests 1–24

This is a much more extensive and complete series of tests on the text dealt with than the one by Bishop and Robinson previously described. The so-called assignments are in reality tests consisting of twenty elements in various forms, such as direct-recall, true-false, completion, matching, multiple-answer, and so forth. Each covers a few paragraphs in the text. The assignments are grouped under twenty main heads, under each of which are

three subheads containing in most cases two, but in some, one or three assignments each. The topical tests are similar to the assignment tests in construction, but deal with particular topics through the whole period of history. Examples of the topics are: agriculture, currency and finance, labor, and political parties. The tests are in twelve pairs, there being two on each topic. As should be evident from the title and description, these are not standardized tests but are for instructional purposes.

Ginn and Company. 96¢ per set.

STANDARDIZED OBJECTIVE EXERCISES IN AMERICAN HISTORY
W. T. Betts (1929)
Exercises I–XX; A, B of each

This series is similar to Stormzand's and other practice tests rather than to ordinary standardized tests. Apparently the twenty A exercises are intended for the first semester and the B ones for the second. Each calls for twenty responses to some one of seven types of exercises. It is stated that the content was "determined largely by the conceptions of important information included in leading modern American histories. No time limits are stated, but few of the tests should require much if any more than five minutes of actual working time.

Southern Publishing Company. Complete set 40¢; 32¢ per 30.

READING SCALES IN HISTORY
M. J. Van Wagenen (1921)
Forms A, B

A description of these scales will be found on page 143, along with that of the whole series of which they are a part.

II. European and World History

Although, as has already been stated, the number of tests in other divisions of history than American is small as compared with the number therein, yet their average merit is much higher

Practically all of the few that are now commercially available appear to deserve mention in this section. About as many more as are mentioned here have been published but can no longer be procured, at least in larger numbers than sample copies. As will be seen from the tests described here, there is none in ancient history except as it is included in the more general topic of world history. Moreover, there are no standardized tests dealing with the history of any particular foreign countries, even though courses in English history and occasionally in that of other countries are sometimes offered in the secondary school.

Diagnostic Test in Modern European History
C. G. Vannest (1921)

In this, the first standardized test covering modern European history, there are five parts dealing with time sense, place sense, evaluation of facts, thought, and information. The first requires the arrangement of the names of persons, events, and periods in chronological order. The second, in multiple-answer form, calls for indicating the nationality of historical characters and the countries in which battles and other events took place. The test on evaluation of facts calls for the selection of the most important events out of several lists. In that on thought are ten paragraphs giving certain historical and related information, each followed by one, two, or three questions supposed to require pupils to reason about the information in the paragraph in order to give correct answers. The last subtest contains six exercises calling for various items of historical information dealing with characters, events, boundaries, and so forth. One rather unusual feature is that in the first three subtests pupils are instructed to mark each name or item that they have not had in their history work. The test, as its name implies, is intended to be diagnostic, and as Vannest suggests "to find out as much about the teacher as about the pupil." It is based entirely upon Part II of Robinson and Beard's *Outlines of European History*, the text used in the high schools of Indiana at the time this test was constructed. No time limit is given. Apparently, however, the test requires about an ordinary high-school period. The following median scores are reported:

Test	*A*	*B*	*C*	*D*	*E*
Median	15	15	18	40	12

Indiana University Book Store. 10¢ per copy, 22¢ per 11, 75¢ per 50.

Reference: Vannest, C. G. "Diagnostic Test in Modern European History," *Bulletin of the Extension Division, Indiana University*, Vol. 6, No. 12. Bloomington: Indiana University, 1921, p. 59–68.

Tests in Mediaeval and Modern History

C. A. Gregory and A. D. Owens (1926)

Forms A, B

The seven parts of this test are as follows:

1. Miscellaneous facts and dates—forty single-answer elements
2. Association of events—twenty multiple-answer
3. Historical vocabulary—ten multiple-answer
4. Sequence of events—twelve multiple-answer
5. Cause and effect—twelve multiple-answer
6. History of the Middle Ages—ten multiple-answer
7. Modern history—thirty multiple-answer

It has comparatively little diagnostic value, but yields general measures. The time covered is from the fall of Rome to the present. About one period is required to give it. The norm is given as sixty-nine and the coefficient of reliability as about .90.

Bureau of Administrative Research. Specimen set 10¢; 4½¢ per copy, $4.00 per 100.

Test in World History

M. W. Sloyer (1928)

This consists of two portions of which the first is merely a single compound matching exercise, whereas the second contains ninety completion statements with from one up to fifteen blanks in each. The period covered is from the earliest times down to the present. All items included were common to at least three textbooks. No time limit is set, but Sloyer states that an hour is sufficient for all pupils.

Palmer Company. 20¢ per copy, 50 per cent discount on quantities.

American Council European History Test

H. J. Carman, W. C. Langsam, and B. D. Wood (1928)

Forms A, B

In the first of the four parts are seventy true-false statements, in the second ten matching exercises with five elements in each, in the third forty-five multiple-answer exercises, and in the fourth twenty completion statements. The period covered is approximately the last five centuries. The actual working time is ninety minutes. This is undoubtedly much the best available test over the period covered, although its reliability is not very high.

$$r' = .82,\ \text{P.E.}_{\text{meas.}} = 8,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .12,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .29.$$

Norms for Form A for six hundred high school pupils at the end of the year and three hundred college entrants are:

	Percentile						
	5	*10*	*25*	*50*	*75*	*90*	*95*
High school	37	44	60	86	110	131	141
College	12	22	34	47	65	82	93

Form B appears to be about fifteen points more difficult than Form A.

World Book Company. Specimen set 25¢; $1.50 per 25.

Modern European History Test

L. J. Ragatz (1929)

Form 1

This test, one of the George Washington University series, is apparently intended for both high-school and college use. It consists of three parts. In the first are twenty-five multiple-answer exercises, in the second one hundred thirty true-false statements, and in the third two matching exercises of ten items

each. The test covers the time from the Middle Ages to the present. It deals with dates, characters, events, and other phases of history. The time allowance is forty-five minutes. The chief adverse criticism of this test is that so much of it is in true-false form.

Center for Psychological Service. 7¢ per copy, $5.00 per 100.

III. Civics

In general the content dealt with by civics tests is wider than is denoted by the narrow meaning of the term. Half or more of the available tests in this field are not at all limited to the operations of government and the participation of citizens therein, but include items drawn from what might be called "citizenship" in a rather broad sense. This, of course, is in accord with the trend of the times in regard to the subject matter of civics courses. In dealing with this material, however, the test maker is faced with a somewhat unusual difficulty. It is much more difficult for him to find definite significant facts and information in sufficient quantity for the construction of two or more forms of a test of adequate length. It is, of course, in place to test knowledge of such facts as eligibility requirements for office, length of term served, number of office holders of particular types, and so on, but these are not really the most important desired outcomes of the subject. Habits, attitudes, and ideals are supposed to receive much emphasis, but are relatively difficult to test.

In addition to the measuring instruments of the type included in this section there are numerous others, called citizenship tests or scales or by some similar term, that seem to be better classified along with personality and character measures than here. Many of these deal with what might be called school citizenship, that is, with such traits as studiousness, perseverance, punctuality, obedience to regulations, coöperativeness, cheerfulness, and so forth, and have no connection at all with civics in its usual meaning. Others approach it somewhat more nearly in that the qualities dealt with are more largely those of daily life, and hence those of ordinary good citizenship. None of these, however, are

designed to be predominantly measures of subject matter ordinarily contained in civics texts and taught in civics courses, and therefore do not seem to belong here.

Civics Test

A. W. Brown and Clifford Woody (1926)

Forms A, B

This test originally appeared as the Michigan Civics Test, in which form much of the work of standardization was carried on previous to the date given above. It has three parts, of which the first contains forty multiple-answer exercises dealing with the definition of civic terms and the second eighty yes-no questions. The third contains two sections. The first presents a fairly detailed list of qualifications of two suppositious candidates for a particular office and asks that the better of the two candidates be indicated, and the qualities or characteristics that determine the choice marked. In the second are eight paragraphs dealing with situations of civic interest, or concerned with an individual's civic duties. Below each five possible answers or courses of action are stated, of which the best one is to be selected. The exercises included are based upon subject matter common to at least five out of nine of the most widely used textbooks in civics. Civic information, habits, thinking, ideals, attitudes, and appreciations are dealt with. The material was selected especially for senior high-school use, but is fairly well adapted to the junior high school also. The time required is thirty-five minutes, sufficient that speed should not be a major factor in determining scores. Norms for pupils who have completed a year's work in community civics based on about eight hundred senior and four hundred junior high-school pupils are given as follows:

	Percentile				
	10	*25*	*50*	*75*	*90*
Senior high	99	108	118	126	132
Junior high	77	88	99	110	118

World Book Company. Specimen set 20¢; $1.30 per 25.

Tests on the Federal Constitution

R. G. Patterson (1927)

Forms A, B, C

Although these tests have apparently not received very wide use, and only very tentative norms have been announced, they appear to merit mention since they cover the Federal Constitution more thoroughly than any other tests so far published. The so-called forms are really three different tests. The first is a fundamental fact test, consisting of thirty direct questions, some of which consist of several parts, calling for brief answers. Form B is a completion test containing two sections of the Constitution with fifty words omitted from each. Form C is called a mathematical test. It contains thirty-one questions, some with several parts, which call for numerical answers of various sorts such as the lengths of terms and age and residence qualifications of various officers, the numbers of certain amendments, and so forth. These tests, especially the second, measure word-for-word knowledge of the Constitution rather closely. It appears somewhat doubtful if high-school pupils should devote enough time to the Constitution to be able to make very good scores on such tests as these. The time required appears to be about twenty-five minutes for Tests A and C, and forty for B. High-school medians based on a small number of cases are 18, 29, and 24 for the three tests in order.

Palmer Company. Sample set 20¢; 2¢ per copy, $1.80 per 100.

Tests in Civic Attitudes and in Civic Information

H. C. Hill (1927)

Each test consists of twenty multiple-answer exercises with four suggested responses. Both deal with civics in its broad sense as including matters other than those particularly pertaining to government. In order to select the items included, Hill studied literature in the field of social science, examined courses of study, secured the judgments of leaders of research and of over two hundred teachers, and tried out the items extensively on junior and senior high-school pupils. There are no time limits. Both the tests are much too short for individual diagnosis. A study of the

test on attitudes based on a small number of cases yields the following:

$$r = .60,\ \text{P.E.}_{\text{meas.}} = 1,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .05,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .42.$$

Medians based on seven thousand pupils are:

Grade	*VII*	*VIII*	*IX*	*X*	*XI*	*XII*
Information	8.4	10.8	12.3	12.9	14.3	15.8
Attitudes	13.0	14.5	15.5	16.4	16.9	17.7

It is found that on the information test, boys score somewhat better than girls, the difference averaging about .5 in the elementary school, and 1.0 in the high school. On the attitudes test there is perhaps a slight tendency for the girls' scores to be higher, but the difference is not great enough to be significant.

Public School Publishing Company. Sample set 10¢; $1.00 per 100.

Reference: Fowler, O. F. "The Civic Attitudes of High-School Sophomores," *School Review*, 36:25–37, January, 1928.

A Test in Civic Action

H. C. Hill and H. E. Wilson (1928)

This is similar in form to the Hill Civic Information and Attitudes Tests, and may be considered as the third of the same series. The items included were chosen upon practically the same bases as those in the others. The test states twenty civic situations and asks that the best one of four suggested courses of action in each be indicated. Norms for the end of the year based on a total of over twelve hundred scores are given as follows:

Grade	*VII*	*VIII*	*IX*	*X*	*XI*	*XII*
Median	8.9	10.8	13.9	15.4	16.2	16.9

In all grades except the ninth boys' median scores are somewhat higher than those of girls.

Public School Publishing Company. Sample set 8¢; $1.25 per 100.

A Test of American Civics and Government

J. C. Almack (1928)

Forms I, II

This test is announced as being intended for junior and senior high schools, normal schools, and junior colleges. Each form contains three parts, the first dealing with civic information, the second with civic judgment and policy, and the third with civic terms. There is a total of seventy-six items in multiple-answer form, with only three suggested answers to each. The author states that the exercises were based upon recent textbooks and courses of study. The time limit is twenty-five minutes.

$$r = .91,\ \text{P.E.}_{\text{meas.}} = 4,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .08,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .20.$$

Norms, apparently for the end of the course, are given for about three hundred persons at each level.

	Percentile						
	5	*10*	*25*	*50*	*75*	*90*	*95*
Junior college and normal	46	49	53	58	63	67	70
Senior high	36	46	50	57	63	67	69
Junior high	22	27	34	42	49	56	59

The test as a whole appears to have at least two points of weakness. The use of only three suggested answers instead of four or five is undesirable since it increases considerably the chances of guessing correctly. In general the material dealt with, especially in the first part, has to do with federal government, although there are a few exercises dealing with state and local government and other civic or semi-civic matters. This fact is hardly in accord with the recent trend of emphasis in civics courses. It would perhaps be better to label this as merely a test of American government, leaving out the word civics.

Bureau of Administrative Research. Sample set 20¢; 2½¢ per copy, $2.00 per 100.

Civics Test
W. H. and Virginia N. Burton (1928)
Forms A, B

This test is primarily intended for use in Grades V–IX, but apparently might well be used at the beginning of a senior high-school civics course in order to determine the knowledge possessed by those beginning it, and perhaps also later. Each form consists of sixty multiple-answer elements dealing with civics information, using the term in a broad sense. The elements are stated in question form, and the three suggested answers are definitions or explanations of certain expressions, such as jury, federal, speculator, employee, injunction, market value, and so forth. Indeed, so many of the terms dealt with are outside the field of civics in the narrowest sense that the test might well be given some other name. They were selected, however, on the basis of an analysis of six textbooks in civics and community civics, designed for use in Grades VI–X, and three for the lower grades, twenty recent state courses of study and Reinoehl's analysis of thirty-five state courses of study.[8] A list of 187 basic terms so compiled was checked against one of important civic problems indicated by various authorities, and also with thirty consecutive issues of a metropolitan daily. Ninety-six of the items were given to several hundred pupils, about one-half of the latter interviewed individually to determine whether or not they really possessed the information indicated by their answers, and from the results the two forms were constructed. Pupils are given twenty minutes in which to complete the test, which should be easily sufficient for sixty exercises of the type employed. A coefficient of reliability of about .80 has been obtained.

The norms given by the author are rather unusual in that they were made by pupils who had not studied a textbook in civics, but had had only "the usual work with clippings and current event papers." For native born pupils of good economic status the following mean scores are given for the end of the semester. These are based upon groups of from four hundred to one thousand in each half grade.

[8] Reinoehl, C. M. "Analytic Survey of State Courses of Study for Rural Elementary Schools," *U. S. Bureau of Education Bulletin*, 1922, No. 42. Washington: Government Printing Office, 1923. 116 p.

Grade	*7B*	*7A*	*8B*	*8A*	*9B*
Median	31	33	36	38	40

World Book Company. Specimen set 20¢; $1.20 per 25.

Reference: Burton, W. H. "A Contribution to the Technique of Constructing 'Best-Answer' Tests," *Elementary School Journal*, 25:762–70, June, 1925.

American Council Civics and Government Test

R. D. Leigh, J. D. McGoldrick, P. H. Odegard, and B. D. Wood

(1928)

Forms A, B

The first of the four parts contains one hundred eight true-false statements; the second, thirteen matching exercises in each of which five definitions, explanations, or portions of statements are to be matched with the proper terms or first portions; the third, twenty-four multiple-answer elements; the fourth, twenty-three single-answer or completion elements. This test is probably the most comprehensive and complete measure of this subject available at present. As is true of the other American Council Tests, so this is intended for both high schools and colleges. Though more items deal with the federal government than with any other general topic, yet a fairly large number are concerned with local government and matters of general civic interest. Indeed, some of them could hardly be strictly classed under the subject of civics, but they are undoubtedly matters with which well-informed citizens should be familiar. The time is ninety minutes. A tentative median score of eighty for five hundred students from thirty-five different high schools has been announced.

World Book Company. Specimen set 25¢; $1.50 per 25.

IV. Miscellaneous

Although in many high schools, especially the larger ones, other social studies than history and civics are offered, very little

has been done along the line of constructing standardized tests to measure their outcomes. The only one of this sort that seems to possess any merit worth mentioning is in the subject of economics. Nothing at all appears to exist for sociology, for what is sometimes called unified social science, or for any other similar course. There is, however, another test than the one in economics which seems to belong in this section. This, as will be apparent from its description later, is not intended to be given at the completion of a course to measure achievement therein, but rather at the beginning to provide an inventory of the pupils' knowledge of history and other social studies.

American Council Economics Test

Horace Taylor, T. N. Barrows, and B. D. Wood (1928)

Forms A, B

Each form is made up of three parts. In the first are ninety true-false statements; in the second twelve matching exercises each containing five items in one column and eight in the other; and in the third twenty-five multiple-answer elements. Pupils' working time is ninety minutes. This test is intended for use in both high school and college, and on the whole will probably receive more use in the latter because many high schools as yet give no course in this subject. Average reliability data are:

$$r = .83,\ \text{P.E.}_{\text{meas.}} = 7,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .07,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .28.$$

Norms for Form A based on one thousand high-school but only a few over one hundred college students are given below. Form B is eight points harder.

	Percentile						
	5	*10*	*25*	*50*	*75*	*90*	*95*
College	64	78	94	109	127	137	142
High school	51	60	76	94	110	125	133

World Book Company. Specimen set, 25¢; $1.30 per 25.

Harvard Background Test in Social Studies

Tyler Kepner (1924)

Forms A, B

This test consists of seven exercises on association of characters and events, literary background, geography, historical vocabulary, social and economic vocabulary, order of events, and dates of events. Multiple-answer, completion, matching, and rearrangement types of exercises are employed. There is no definite time limit, but it is stated that forty minutes are ample for high-school freshmen. The primary function of these tests is to diagnose the factual background in the social studies possessed by pupils at the beginning of a course. More of the items called for deal with history than any other field, but in addition civics, economics, geography, and sociology are included. Twenty-one representative textbooks published from 1919 to 1922, inclusive, all existing tests in history, and certain other sources furnished the material included. The items selected were analyzed critically by teachers and tried out with pupils, a number being dropped at each step. It is not apparent from an examination of the test that it is so constructed as to differ materially from 'ordinary' achievement tests except perhaps in including a somewhat wider range of material than is found in a single such test. The suggested answers to several of the multiple-answer exercises are not well chosen, either containing two that are correct or somethat are doubtful. The following reliability figures have been reported:

$$r = .79,\ \text{P.E.}_{\text{meas.}} = 2.5,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .05,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .31.$$

The correlations between this test and ordinary history tests range from about .50 to .80, which is at least as high as the average intercorrelation of standardized history tests.

Ginn and Company. $1.00 per 30.

Reference: Kepner, Tyler. "An Aspect of History Testing," *Historical Outlook*, 15:414–16, December, 1924.

BIBLIOGRAPHY

I. American History

Bagley, W. C. and Rugg, H. O. "The Content of American History as Taught in the Seventh and Eighth Grades," *University of Illinois Bulletin*, Vol. 13, No. 52, School of Education Bulletin No. 16. Urbana: University of Illinois, 1916. 59 p.

Bell, J. C. and McCollum, D. F. "A Study of the Attainments of Pupils in United States History," *Journal of Educational Psychology*, 8:257–74, May, 1917.

Brinkley, S. G. "Values of New Type Examinations in the High School with Special Reference to History," *Teachers College, Columbia University, Contributions to Education*, No. 161. New York: Bureau of Publications, Teachers College, Columbia University, 1924. 121 p.

Elston, Bertha. "Improving the Teaching of History in the High School through the Use of Tests," *Historical Outlook*, 14: 300–5, November, 1923.

Gibson, O. H. "Existing Standard Tests in History," *Historical Outlook*, 12:324–26, December, 1921.

Gold, M. S. "Testing Vocabulary in History," *Historical Outlook*, 17:285–91, October, 1926.

Harlan, C. L. "Educational Measurement in the Field of History," *Journal of Educational Research*, 2:849–53, December, 1920.

Henmon, V. A. C. "Some Limitations of Educational Tests," *Journal of Educational Research*, 7:185–98, March, 1923.

Johnson, H. *Teaching of History in Elementary and Secondary Schools*. New York: The Macmillan Company, 1915. 497 p.

Kepner, P. T. "A Survey of the Test Movement in History," *Journal of Educational Research*, 7:309–25, April, 1923.

Odell, C. W. "The Use of Scales for Rating Pupils' Answers to Thought Questions," *University of Illinois Bulletin*, Vol. 26, No. 36, Bureau of Educational Research Bulletin No. 46. Urbana: University of Illinois, 1929. 34 p.

Ruch, G. M., and Stoddard, G. D. *Tests and Measurements in High School Instruction*. Yonkers, New York: World Book Company, 1927, p. 177–82.

Ruch, G. M., et al. *Objective Examination Methods in the Social Studies*. Chicago: Scott, Foresman and Company, 1926. 116 p.

Rugg, E. U. "Character and Value of Standardized Tests in History," *School Review*, 27:757–71, December, 1919.

Sackett, L. W. "A Scale in United States History," *Journal of Educational Psychology*, 10:345–48, September, 1919.

Symonds, P. M. *Measurement in Secondary Education*. New York: The Macmillan Company, 1927, p. 162–76.

Tryon, R. M. "Standard and New Type Tests in the Social Studies," *Historical Outlook*, 18:172–78, April, 1927.

"Objective Achievement Tests Constructed and Used in St. Louis," (*St. Louis*) *Public School Messenger*, 25:148–75, November 30, 1927.

II. European and World History

Hefley, Sue. "Test on World History from Earliest Times through the Reformation," *Historical Outlook*, 20:27–29, January, 1929.

Kepner, P. T. "A Survey of the Test Movement in History," *Journal of Educational Research*, 7:309–25, April, 1923.

Rugg, E. U. "Character and Value of Standardized Tests in History," *School Review*, 27:757–71, December, 1919.

Sackett, L. W. "A Scale in Ancient History," *Journal of Educational Psychology*, 8:284–93, May, 1917.

Tryon, R. M. "Standard and New Type Tests in the Social Studies," *Historical Outlook*, 18:172–78, April, 1927.

III. Civics

Barr, A. S. "Measurement in Civics," *Historical Outlook*, 17:24–26, January, 1926.

Chassell, C. F. and Chassell, E. B. "A Test and Teaching Device in Citizenship for Use with Junior High School Pupils," *Educational Administration and Supervision*, 10: 7–29, January, 1924.

Chassell, C. F. and Upton, S. M. "A Scale for Measuring the Importance of Habits of Good Citizenship," *Teachers College Record*, 20:36–65, January, 1919.

Chassell, C. F., Upton, S. M., and Chassell, L. M. "Short Scales for Measuring Habits of Good Citizenship," *Teachers College Record*, 23:52–79, January, 1922.

Hill, E. L. "A Citizenship Rating Scale," *Education*, 47: 362–71, February, 1927.

Odell, C. W. "The Use of Scales for Rating Pupils' Answers to Thought Questions," *University of Illinois Bulletin*, Vol. 26, No. 36. Bureau of Educational Research Bulletin No. 46. Urbana: University of Illinois, 1929. 34 p.

Ruch, G. M., and Stoddard, G. D. *Tests and Measurements in High School Instruction*. Yonkers, New York: World Book Company, 1927, p. 182.

Stack, H. J. "Standardized Tests in Community and Economic Civics," *Historical Outlook*, 18:166–72, April, 1927.

Symonds, P. M. *Measurement in Secondary Education*. New York: The Macmillan Company, 1927, p. 179–82.

Tryon, R. M. "Standard and New Type Tests in the Social Studies," *Historical Outlook*, 18:172–78, April, 1927.

IV. Miscellaneous

Symonds, P. M. *Measurement in Secondary Education*. New York: The Macmillan Company, 1927, p. 176–79.

CHAPTER X

MANUAL ARTS AND HOME ECONOMICS

Introduction.—It is more difficult to measure adequately the desired outcomes of instruction in the subjects dealt with in this chapter than in most of those considered in previous chapters. The reason is that the desired outcomes are largely skills, manual or otherwise, which it is difficult to test satisfactorily by means of pencil and paper tests. It would be possible to specify standard conditions under which pupils might carry out actual projects in these fields, such as planing a board, making a joint, or even constructing an entire article in manual training, preparing some article of food or doing a specified bit of sewing in home economics. Indeed, a number of tests of this sort have been employed, but they are practically all intended for use in industry or elsewhere rather than in school. The objection to them is that they are expensive of time, frequently requiring individual administration, and that it is difficult to secure absolutely uniform conditions and objective rating of the product. Measurement, therefore, has been largely limited to the testing of information and knowledge of what to do and has not dealt with the actual ability to do it. It is true that the two phases of the work correlate positively, since it is impossible for one to do a good piece of work if he does not know how to do it. On the other hand, he may know how without possessing the proper technical vocabulary that will enable him to respond correctly to exercises concerning it; and still more likely he may know how and be able to respond to the exercises correctly, but not possess the necessary skill to put his knowledge into practice.

I. Manual Arts

The term "manual arts" is used here to include such courses as are commonly given in high schools under the name of manual arts, industrial arts, shop work, or some other similar title.

Especially in rather large city high schools such courses exist in wide variety, including many types of woodwork, metal work, mechanics, and so forth. As will be seen from the next few pages, the tests actually available and recommended for school use are very limited both in number and scope. There are none at all for most of the courses offered. This statement does not mean that no tests at all along these lines exist, since many industrial concerns and others have devised and employed such tests. They have not, however, received any considerable use in school, and so will not be included here. It appears in place, however, to refer to the tests of this sort constructed for use in the United States army during and since the World War [1] inasmuch as these constitute probably the most complete series of the sort. These tests are designed to determine the amounts of ability possessed by recruits in the various vocations dealt with. They cover such different occupations as those of blacksmith, boiler maker, carpenter, foundryman, plumber, printer, sheet metal worker, welder, and so on. In many instances there are several tests for subdivisions of the same general occupation; for example, there are tests for the following varieties of blacksmiths: forger and hammersmith, drop forger, horse shoer, and general blacksmith. These tests are of three varieties: individual oral, group written, and performance tests, that is, tests of the sort referred to above in which the testees are required to perform an actual piece of work.

One difficulty in the way of constructing manual arts tests for use in school is that the courses offered in this subject vary greatly. The amounts of time devoted to the subject differ considerably, some schools giving two or three periods a week, others five periods or even more; in some the periods are of ordinary length, in others perhaps sixty minutes, and in others eighty or ninety. Moreover, the content of courses, especially the objects actually constructed, depends more largely upon the desires of the instructor and pupils than is true of most book subjects. For these reasons it is difficult to construct tests to fit the courses taught in the different schools, and even in the same school from year to year.

[1] Chapman, J. C. and Chapman, Daisy. *Trade Tests.* New York: Henry Holt and Company, 1921. 435 p.

Shop Tests

W. L. Hunter (1927)

Tests W–1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14; MD–1, 2, 3, 4; MS–1, 2, 3, 4, 5; E–1, 2,; AM–1, 2; P–1; G–1, 2, 3, 4

This is by far the most elaborate series of tests in this subject at all suitable for school use. In the titles of the tests "W" stands for woodworking, "MD" for mechanical drawing, "MS" for machine shop, "E" for electricity, "AM" for auto mechanics, "P" for printing, and "G" apparently for general, as the G tests include one in shop English and three in shop mathematics. Each test consists of a single sheet calling for twenty-five responses in completion, alternative, multiple-answer, single-answer, or some other form. Many of the tests contain drawings and figures upon which the exercises are based. The time limits for more than half of the tests are five minutes and for the others range up to twenty. The tests appear to provide fairly satisfactory measures of the topics dealt with, and the series as a whole may be considered diagnostic since each test deals with a particular portion or phase of the whole subject. For example, the first of the woodworking tests deals with tools, the second with fastenings, the fifth with the use of rules or scales, the tenth with wood and lumber, and so on. A few of the tests are general rather than specific.

Manual Arts Press. 25¢ per 25.

Industrial-Arts Tests

H. B. Nash and R. R. Van Duzee (1927)

Test I—Woodworking; Scales A, B

Scale A, Technical and Related Information, consists of sixty true-false statements, forty-two multiple-answer ones, a number of matching exercises, a completion test involving the reading of a drawing, and a test that involves the correction and completion of figures. Scale B, Performance, contains working and pictorial drawings and directions for preparing a block of wood. This is to be done in such a way as to require the use of a number of different tools. Forty minutes are allowed for Scale A; for B

apparently there is no time limit. The exercises in A are based on material common to representative courses of study, textbooks, surveys, and special references and books on trade analysis.

$$r = .86,\ \text{P.E.}_{\text{meas.}} = .05,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .08,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .25.$$

Norms for Scale A are given as follows, based on the time the subject has been studied:

Semester	Percentile				
	10	25	50	75	90
Training school	83	100	110	122	130
IX and up	59	70	82	100	119
VIII	55	65	78	95	112
VII	48	60	75	93	107
VI	44	55	71	88	102
V	41	50	65	75	89
IV	37	44	56	70	85
III	32	41	52	65	76
II	26	36	46	56	73
I	25	33	43	54	70

These appear to be derived from pupils who have been in junior high school for the first six semesters and have given only from about an eighth to two-fifths as much time per semester to the study of the subject as is common in senior high-school classes.

Bruce Publishing Company. Specimen set 65¢; $1.10 per 25; manual 50¢.

Industrial Arts Tests

G. K. Wells and M. L. Laubach (1928)

Woodworking, Printing, Machine-Shop, Mechanical Drawing; Form A of each

Each test except the last consists of one hundred true-false statements and twenty-five completion statements or multiple-choice exercises covering in a general way the subject dealt with. That on mechanical drawing has only the hundred true-false statements. No information is available as to the selection of content, but apparently it was based upon what is more or less

commonly taught. Twenty-five minutes are allowed for each of the first two tests, twenty for the third, and thirty for the last. Probably the most desirable feature of the tests is that so many of the responses called for are to be given in true-false form. Tentative medians based on about 1,000 cases are as follows:

	Semester			
	I	*II*	*III*	*IV*
Woodworking	51	54	65	66
Printing	47	50	58	62
Machine shop	49	57	66	75
Mechanical drawing	35	43	50	54

Manual Arts Press. 25¢ per 5.

Home Mechanics Test

L. V. Newkirk and G. D. Stoddard (1928)

Forms A, B

Part 1 contains thirty-two or thirty-three exercises in each of which the steps in the procedure of carrying out a common job in home mechanics are given in confused order and must be correctly rearranged. In Part 2 are representations of the parts involved in three or four electrical jobs such as connecting dry cells in series or in parallel, hooking up the aerial of a radio, and so forth, with general statements of what is to be done. Pupils are to draw lines to indicate the proper connections or circuits. On the basis of a study which checked jobs according to their social utility, surveyed the actual teaching content of seventy-five schools, analyzed courses of study and widely used commercial job sets, and selected jobs with procedures representative of classes of jobs rather than of single ones, seventy-two high ranking jobs in home mechanics were selected and arranged into two equivalent forms. Forty minutes are allowed pupils for taking the test. Two scores are provided for, one a job score based upon the number of jobs completely right and the second a point score based upon the number of procedures correctly placed whether or not the other procedures on the same jobs were correct or not. The average reliability of the test is about as follows:

Job scores:

$$r = .57,\ \text{P.E.}_{\text{meas.}} = 1.3,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .25,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .44.$$

Point scores:

$$r = .87,\ \text{P.E.}_{\text{meas.}} = 6.5,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .09,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .24.$$

Thus it appears that when point scores are employed the reliability is fairly satisfactory, but that when job scores are used it is decidedly low. Norms for May testing based on only about four hundred cases are approximately as follows:

	First quartile	*Median*	*Third quartile*
Job score	3	5	7
Point score	57	75	96

Bureau of Educational Research and Service. $4.00 per 100; manual 10¢.

II. Mechanical Drawing

Although the subject of mechanical drawing is closely connected with manual arts, it is frequently taught as a separate course and it has, therefore, seemed desirable to deal with its measurement in a separate section. There are at least three possible methods of measuring ability in this subject. One is to employ a rating scale similar in purpose and use to those in handwriting, composition, and drawing, by means of which specimens of pupils' work are scored. Another is to test information concerning the subject. The former of these methods is open to the objection that it is not entirely objective, and the latter that it does not deal with the chief desired outcome of instruction. A third possible method consists in having pupils make certain drawings or portions of drawings of such a nature that the construction can rather definitely be marked either right or wrong. This, on the whole, appears preferable to either of the other two, but can hardly be extended to test all that it is desired to measure.

Grading Chart for Mechanical Drawing

P. M. Spink

This chart is similar in form to a handwriting scale. It presents a number of specimens at each of six degrees of merit. Most of these specimens are of mechanical lettering, but there are also a few simple figures. The five highest degrees of merit compose a scale for high-school use and the five lowest make one for the elementary school. Each scale has ratings from fifty-five to ninety-five, inclusive, by tens. Since four of each set of five are common, the ratings on these four are in each case 10 per cent lower for high-school than for elementary-school pupils. No standardized directions are given for securing samples of work, and the suggestions as to the use of the chart are very brief. Furthermore there is no information as to how the samples included were secured or their merit determined, nor are any norms given. Apparently the objectivity of scores according to this chart is about the same as according to a handwriting scale.

Safety Electric Heater Company. 75¢ per copy, $3.00 per 6.

Mechanical Drawing Test

D. W. Castle (1928)

In the first of the five subtests are three exercises each of which presents a top and a side view of the same object. There are several numbers on one view and several letters at corresponding points on the other. Those taking the test are to match the numbers and letters. The second subtest deals with dimensions. Several figures are given, with certain dimensions of each, and others are to be determined from those given. The third deals with the knowledge of fourteen geometrical terms, presented in multiple-answer definition and completion form. The fourth subtest, on pencil technic, calls for a complete pencil copy of a given drawing. The last, on linking technic, requires that a drawing printed very faintly so as to imitate pencil work be inked. No information concerning the source of content is given other than that geometrical drawing was analyzed into certain divisions and appropriate testing devices formulated. The total working time for the test is forty-one minutes. The

scoring of the first three parts is perfectly objective, but that of the last two is not, being done according to the opinion of the scorer. However, six points on which each drawing is to be scored are stated and also the number of points allowed on each. Furthermore, letter rating scales are provided for use in connection with the drawings. Thus subjectivity is decidedly reduced.

Manual Arts Press. 25¢ per 5.

Mechanical Drawing Test
G. K. Wells and M. L. Laubach (1929)
Form A

This is described along with the series of industrial arts tests by the same authors in Section I of this chapter.

A Standard Test in Fundamental Mechanical Drawing
A. J. Badger (1929)
Tests 1, 2, 3

The first of these tests deals with the use of tools, line work, dimensioning and lettering; the second with projection, including sections and auxiliary views; and the third with pictorial drawing, isometric, cabinet, and oblique. The first contains seventy-five test elements, the second forty, and the third thirty. All are in multiple-answer form, but several varieties thereof are employed. A number of exercises require the marking or labelling of figures. The author states that "these are tests of what the pupil knows about the phases of drawing covered rather than a test of his drawing ability measured in terms of neatness, accuracy, lettering, and so forth." Directions provide no time limits, but state that the booklets should be collected when all but the slowest two or three pupils have finished. Seemingly each test should easily be completed within twenty or twenty-five minutes at least.

Public School Publishing Company. Sample set 20¢; $2.00 per 25 of all three tests.

III. Home Economics

The rather rapid development of this subject in the last twenty or twenty-five years has been accompanied by a considerable amount of activity in test construction. Most of this has had to do with courses in foods and clothing and comparatively little with such important phases of the subject as general home management, care of the house, budget making, and so forth. This is largely a reflection of the conditions that actually exist with regard to the courses offered, but nevertheless is unfortunate. As will be seen from the measuring instruments described in this section, the rating scale idea has received some use here as well as in mechanical drawing. Most of the tests, however, deal rather with information concerning foods and clothing and their preparation.

Home Economics Information Tests for Girls Completing the Eighth Grade

Department of Household Arts Education, Teachers College, Columbia University (1923)

Sets I, II, III

Although these tests were designed for use at the end of the elementary-school period, there seems to be no good reason why they should not also be valuable for high-school use. The three sets deal with clothing, foods, and other household activities. They are supposed to cover the minimum essentials of these subjects as determined by a study of textbooks and of opinions of university instructors and public-school teachers of the subject. The first set has subtests on textiles, the construction of clothing, care and repair of clothing, and selection in clothing. Set II has subtests on sources of our common foods, food selection, food preservation and storage, laboratory practices, food values and health in meal selection, and food preparation. The subtests in Set III deal with the girl's bedroom, the dining room, dishwashing, care of the kitchen, labor-saving devices, home enjoyment, care of children, and the budget. Each set contains about a hundred multiple-answer elements. All the time needed is to be given, but a class period should suffice. Results

from over a thousand eighth-grade pupils from forty schools are given in terms of per cents of errors. The general average per cent is about twenty-one for the set on clothing, twenty-two for that on foods, and fourteen for that on household activity. Similar figures are also given for each of the subtests.

Bureau of Publications. Sample set 25¢; 15¢ per series.

Reference: Cooley, A. M. and Reeves, Grace. "Some Investigations Concerning the Use of Certain Home Economics Information Tests," *Teachers College Record*, 24:374–92, September, 1923.

Illinois Food Test

Anna B. Robinson, Adah H. Hess, Georgina Lord, and Mabel Trilling (1924)

This test consists of one hundred multiple-answer elements arranged under fourteen different heads having to do with the selection, preparation, and serving of food, and so forth. Although it deals primarily with information, some of the exercises are supposed to test judgment also. The authors, a committee appointed by the Illinois Home Economics Association, sent out a large number of the tests in preliminary form and after securing results from some thousands of pupils in over one hundred schools, drew up the final form. Its division into parts, each dealing with a single topic, gives it some diagnostic value, but in general the parts are too short for this to be very great. The test as a whole, however, yields a rather good general measure of the subject tested. It requires thirty-five minutes. The median April score is seventy-five.

Public School Publishing Company. Sample set 10¢; 75¢ per 25.

Reference: Hess, Adah. "Tests in Home Economics," *High School Conference Proceedings*, 1922. Urbana: University of Illinois, 1923, p. 249–54.

Foods Test

Florance B. King and H. F. Clark (1926)

Form A

This is a revision of an earlier test by Miss King. It consists

of sixty multiple-answer elements. The selection of content was based upon an analysis of six of the nine [2] textbooks used by 90 per cent of home-economics teachers throughout the country, and of three published more recently. Also courses of study and articles on home economics were used as sources. The tentative results were discussed with teachers of foods in elementary, secondary, and higher institutions, and tried out with both such teachers and pupils. The working time is thirty minutes. Median scores for April based on about five thousand pupils in all are as follows:

Grade	*VII*	*VIII*	*IX*	*X*	*XI*	*XII*
Median	30	33	36	38	40	44

World Book Company. Specimen set 10¢; $1.00 per 25.

Reference: King, F. B. "A Measuring Scale in Foods," *Bulletin of the Extension Division, Indiana University*, Vol. 7, No. 12. Bloomington: Indiana University, 1922, p. 144–46.

A Test at the Completion of Food Preparation Courses in the Junior High School

Nina Streeter and Mabel Trilling (1927)

This is a short, easily administered test of twenty-five multiple-answer exercises covering the topic indicated. It is based upon textbooks and courses of study in use at the time it was constructed. It is stated that the test involves judgment as well as information, but examination of it indicates that the latter is decidedly predominant. Seventeen minutes are allowed, which should be sufficient for practically all pupils. If one desires a brief test to yield general measures of knowledge of food preparation, this is probably the best test available. It is, however, too short to yield measures of diagnostic value.

Public School Publishing Company. Sample set 10¢; 75¢ per 25.

[2] The three omitted were either out of date or not sufficiently comprehensive.

A Sewing Scale
Katharine Murdoch (1919)

This scale consists of fifteen plates, each of which contains three views of a sewing sampler involving several kinds of hand stitches. The plates are valued at more or less equal steps from 0 up to 16.4. The scale was constructed by securing samplers from over twelve hundred persons, including school pupils, college students, adults, and mental defectives. These were judged, a selection made, then a further judgment and selection, and so on, until 347 judges had participated in the selection of the fifteen specimens pictured in the scale. For separate stitches the average coefficients of reliability between average ratings of one group of ten persons and those of another range from .82 to .95, and for all stitches combined average .96. The intercorrelations between the various stitches run from .26 to .83, and the correlation of each stitch with general ratings based upon all from .62 to .80. The correlation between ratings of sets of two samplers each made by the same pupil range from .40 to .80 on separate stitches and is .72 for all combined. These figures are corrected for attenuation.[8] The uncorrected coefficients are not reported, but are probably from .10 to .20 lower.

Such a scale as this is not as valid as a spelling, drawing, or composition scale because photographs of sewing do not reproduce the sewing itself perfectly. Nevertheless, they have been found to result in standardizing teachers' ratings of pupils' sewing. Correlations of from .36 to .84 have been found with teachers' marks. The average deviation of judgments of girls preparing to be teachers of home economics is slightly greater than one point. The following norms based on an average of about two hundred cases in each grade have been reported:

[8] Attenuation is the reduction in the value of an actually computed coefficient of correlation caused by variable or chance errors in the measures correlated. If the reliability of the measures is known, the computed coefficients can be corrected and, of course, raised, by applying the proper formula so as to indicate the true correlation between the things measured. For a fuller discussion see:

Odell, C. W. *Educational Statistics.* New York: The Century Co., 1925, p. 181–85, or some other text on the same subject.

Grade	VII	VIII	IX	X	College	
					1st yr.	2nd yr.
Norm	6	7.6	9	10.8	11.5	12.7

The time required to secure samplers similar to those in the scale is about forty-five minutes for high-school freshmen and not much more than thirty for college students.

Bureau of Publications. $1.50 per copy.

Reference: Murdoch, Katharine. "The Measurement of Certain Elements of Hand Sewing," *Teachers College, Columbia University, Contributions to Education*, No. 103. New York: Bureau of Publications, Teachers College, Columbia University, 1919. 120 p.

Brown, C. M. "Investigations Concerning the Murdoch Sewing Scale," *Teachers College Record*, 23:459–70, November, 1922.

Analytic Sewing Scale for Measuring Separate Stitches
Katharine Murdoch (1923)

This is in a sense a shortened form of the scale just described. It consists of one double sheet on which are five views of each of the following five stitches: running, combination, back stitch, overcasting, and hemming. The specimens are at values ranging by steps of 2.5 from six to sixteen, inclusive, these values having the same meaning as on the original Murdoch scale. It is also provided that scores may be given halfway between the values in the scale. This scale is intended for measuring each stitch separately, whereas the earlier one was primarily for measuring general merit in sewing. Furthermore, it is designed largely for the use of pupils themselves, whereas the other was chiefly for teachers. The facts that the former scale was rather inconvenient to use, had unequal intervals, and was decidedly expensive, were also influential in causing the construction of this one. It was made by securing a hundred samples of each stitch, having these rated by twenty competent judges, and then selecting five at approximately equal intervals of merit according to values on the former scale. The same norms apply for this as for Murdoch's original scale.

Bureau of Publications. 25¢ per copy; discount on class orders; manual 10¢.

Reference: Murdoch, Katherine. "A New Analytic Sewing Scale," *Teachers College Record*, 23:453–58, November, 1922.

Tests in Comprehension of Patterns
L. Stevenson and Mabel Trilling (1927)

The five subtests deal with recognition of parts of patterns, comprehension of pattern lines and notches, alteration of patterns, and placing the pattern on the material. Each consists of representations of patterns or parts thereof which the pupils are to mark in various ways to indicate the knowledge called for. Twenty-five such exercises are included. The patterns included are sample foundation ones and their selection was based on the content of high-school courses of study. Although the scoring is supposed to be objective, it appears that there is considerable room for subjectivity in the interpretation of pupils' responses. For example, in many cases pupils are to draw a cross upon a certain line or at a certain place, to draw a line showing an alteration, and so forth, and it is sometimes difficult to decide whether the cross or line has been placed correctly or not. Instructions are to allow sufficient time for everyone to finish. Apparently this should not be more than half of an ordinary class period.

Public School Publishing Company. Sample set 15¢; $1.00 per 25.

Clothing Test
Florence D. Frear and W. W. Coxe (1928)

The five parts of this test deal, respectively, with fundamentals of clothing construction, care and repair of clothing, selection of clothing from the standpoint of hygiene and from that of appropriateness, and economics of clothing. The number of responses called for by each varies from ten to fifty, the total being about one hundred fifty. Parts 2 and 4 are in alternative form, the other three in multiple-answer form. The test is designed to measure knowledge of clothing possessed by high-school girls

taking or having completed courses in dressmaking. It deals with the fundamental principles and processes of clothing construction and related subject matter. Ten phases dealt with are specified. It was originally planned to measure the information called for by the New York State Syllabus in this subject. After several revisions on this basis it was tried out in several other states and as a result organized into the present form. A list of correct answers is provided, but it is suggested that in some cases teachers may disagree. Tentative norms for the ends of the first three semesters of clothing instruction in high school are, respectively, fifty-one, sixty, and sixty-eight. Semester gains of only nine and eight on a test of this length make it evident that either comparatively little progress is made after the first semester of instruction or the test does not discriminate satisfactorily. No definite amount of time is prescribed, but it is suggested that papers should be collected when all but 10 per cent have finished, and that this will require about an hour.

Public School Publishing Company. Sample set 15¢; $1.25 per 25.

Achievement Scales in Household Science

May E. Davis (1928)

Scales A, F, L, R, S; Division I of each

The five scales are as follows: A—Food composition and diet planning, F—Marketing and housewifery, L—Care, preservation, and preparation of food, R and S—Comprehensive scales. The scales are announced as being intended for Grades VIII and IX. Each of the first three consists of thirty multiple-answer exercises apparently arranged in order of difficulty. The comprehensive scales contain sixty such elements apiece. Provision is made for correcting the raw scores to secure final scores similar to those used by Van Wagenen in many of his tests. The items included appear to be such that the tests yield fairly good general measures of achievement. They are, however, scarcely long enough to be diagnostic, even though there is a test upon each of several divisions of the subject. The time is forty minutes. Tentative norms are given as follows:

Grade	Scale				
	A	*F*	*L*	*R*	*S*
Low IX	84	86	87	84	84
High VIII	—	—	—	81.5	81.5
Low VIII	—	—	—	80.5	80.5

Ginn and Company. $1.00 per 30.

Technical Information Test for Girls
B. D. Leary and R. R. Dry (1924)

In this test are one hundred multiple-answer statements dealing with various facts in the fields of cooking, homemaking, sewing, care of children, and so forth. They were selected from over six hundred originally made up. It is stated that the purpose is to ascertain technical knowledge in the field of home economics. There is no definite time limit, but it seems that thirty minutes should be ample.

C. H. Stoelting Company. $2.20 per 25, $6.00 per 100.

BIBLIOGRAPHY

I. Manual Arts

Broom, Eustace and Lovelace, L. H. "A Study of a Test in Home Mechanics," *High School Teacher*, 3:262–63, 272, September, 1927.

Chapman, J. C. and Chapman, Daisy. *Trade Tests*. New York: Henry Holt and Company, 1921. 435 p.

Leavitt, F. M. "Standardized Measurement Scales in the Field of the Industrial Arts," *Industrial Arts Magazine*, 8:132–38, April, 1919.

Moss, F. A. "Suggested Tests for Painter," *Public Personnel Studies*, 3: 131–36, April, 1925.

Ward, C. M. and Toops, H. A. "Performance Test of Ability in Using Measuring Tools," *Industrial Education Magazine*, 27:177–80, December, 1925.

"Information and Data Regarding Tests in the Short Answer Form," *Public Personnel Studies*, 6:219–20, October, 1928.

"Objective Achievement Tests Constructed and Used in St. Louis," (*St. Louis*) *Public School Messenger*, 25:91–107, November 30, 1927.

"Suggested Tests for Blacksmith," *Public Personnel Studies*, 6:90–94, April, 1928.

"Suggested Tests for Carpenter," *Public Personnel Studies*, 4:320–24, November, 1926.

"Suggested Tests for Plumber," *Public Personnel Studies*, 3:289–300, October, 1925.

II. Mechanical Drawing

Rugg, H. O. "A Scale for Measuring Free-Hand Lettering," *Journal of Educational Psychology*, 6:25–42, January, 1915.

Starch, Daniel. *Educational Measurements*. New York: The Macmillan Company, 1916. 202 p.

III. Home Economics

Brown, C. M. "Construction and Use of Information Tests in Home Economics," *Journal of Home Economics*, 16:251–56, May, 1924.

————. "What Can Educational Measurements Do for Home Economics," *Journal of Home Economics*, 16:191–96, April, 1924.

Hess, Adah. "Tests in Home Economics," *High School Conference Proceedings*, 1922. Urbana: University of Illinois, 1923, p. 249–54.

McGowan, E. B. "Progress Report on the Petticoat Test," *Journal of Home Economics*, 14:20–23, January, 1922.

O'Brien, F. P. and Giblette, C. T. "A Project Test of Achievement in Sewing," *School Review*, 35:217–21, March, 1927.

Page, E. M. "Tests and Measurements in Sewing," *Chicago Schools Journal*, 3:302–5, June, 1921.

Richmond, J. E. "The Measurement of Informational Facts in Food Study," *Journal of Home Economics*, 17:658–61, November, 1925.

Symonds, P. M. *Measurement in Secondary Education*. New York: The Macmillan Company, 1927, p. 216–21.

Trilling, M. B. et al. "Home Economics in American Schools," *Supplementary Educational Monographs*, Vol. 2, No. 6. Chicago: University of Chicago, 1920. 122 p.

Trilling, M. B. and Hess, Adah. "Informal Tests in Teaching Textiles and Clothing," *Journal of Home Economics*, 13:483–89, October, 1921.

"Suggested Tests for Cooks," *Public Personnel Studies*, 5:125–28, June, 1927.

"Suggested Tests for Dietician," *Public Personnel Studies*, 7:29–35, February, 1929.

"Suggested Tests for Housekeeper," *Public Personnel Studies*, 6:215–18, October, 1928.

CHAPTER XI

MUSIC AND ART

Introduction.—There are at least two reasons why measurement in the field dealt with by this chapter presents more difficulties than is true in the case of many high-school subjects. Music and art are similar to literature in that appreciation is or should be one of the chief objectives of instruction. Also, they are similar to the subjects treated in the last chapter in that one of the principal outcomes is ability to perform activities that cannot well be reduced to writing. There is, of course, a certain amount of definite information that can be tested rather easily, and in addition certain of the products of artistic and freehand drawing ability may be rated by the use of scales and otherwise. These possibilities have been sufficient that the number of standardized tests constructed in music is as great as in a number of other subjects, and in art and drawing a few of merit enough to justify their use have appeared.

The view has been held very generally that ability along musical and artistic lines is much more largely the result of inheritance and less susceptible to development through training than is true in the case of such subjects as mathematics, science, history, and so forth. Furthermore, these so-called "aesthetic" subjects have commonly been thought of as ornamental rather than practical and, therefore, as being of decidedly less importance than many others. It is also true in many cases that it is much more expensive to pursue advanced study along these lines, largely because more individual instruction is necessary, than is the case with most subjects. As a result of these reasons, and perhaps of others also, there seems to be somewhat more interest in prognostic tests in this field than in many others. This is accompanied by the belief that it is probably scarcely worth while to attempt to develop such abilities in children who do not seem to possess sufficient amounts of capacity to profit con-

siderably thereby. Unfortunately some of the most extensive work along this line has not resulted in the publication of available tests and, therefore, will not be referred to here.

I. Music

Many of the tests in this subject are intended for both elementary and high school, and of those that are not, most are for the former rather than the latter. As was suggested above, they deal with knowledge about music, including notation and terminology, musicians and their works, and so forth. Practically nothing has been done by way of testing directly such commonly emphasized objectives as group singing and ability to play in an orchestra, and nothing very satisfactory in the case of musical appreciation. Most of the measurements possible are of achievement, although one outstanding series of tests, the first to be described, is prognostic in its purpose.

Measures of Musical Talent
C. E. Seashore (1915)

This, the earliest and best-known series of standardized tests in music, is prognostic and not for the measurement of actual accomplishment or appreciation. More particularly its function is to determine whether individuals possess sufficient possibility of success as musicians to warrant giving them advanced musical training. The test material consists of six standard double twelve-inch phonograph records dealing with pitch, intensity, time, consonance, memory, and rhythm. The sense of pitch test presents one hundred pairs of tones and requires pupils to discriminate between the higher and lower of each pair. Those in intensity and time are similar. That on consonance presents twenty-five pairs, then repeats the same in reverse order and requires judgments at to which of each pair is heard better. The rhythm test consists of fifty rhythmic patterns, each pair either being alike or differing in time or intensity or both. The tonal memory test likewise has fifty items, ten presenting two-tone patterns, ten three-tone patterns, and so on up to six-tone patterns. Each pattern is repeated with one tone changed and this is to be identified. About ten to fifteen minutes are needed for each record.

It appears that these tests fulfill their function fairly well, at least in a negative fashion. In other words, persons who make low scores have little chance of becoming successful musical performers. High scores, however, do not insure musical success. It should also be understood that low scores do not indicate that those making them should not receive at least such training in music as is commonly given in the public schools, and should not be able to enjoy and appreciate music even if they cannot be solo performers. The average reliabilities reported for the several tests vary within these limits:

$$r = .39\text{–}.72,\ \text{P.E.}_{\text{meas.}} = 3\text{–}6,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .04\text{–}.09,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .36\text{–}52.$$

Intercorrelations ranging from .20 to .78 have been found among scores on the different tests. Combined scores on all the tests correlate about .30 with marks in applied music and .50 with those in music theory. Norms in terms of per cents correct are given as follows:

Percentile	*5*	*10*	*25*	*50*	*75*	*90*	*95*
Pitch							
Eighth grade	54	57	67	77	82	86	88
Adults	61	66	75	81	85	88	90
Intensity							
Eighth grade	67	71	77	83	87	91	93
Adults	74	78	84	89	92	95	96
Time							
Eighth grade	57	61	65	71	76	80	82
Adults	65	69	73	78	82	86	88
Consonance							
Eighth grade	52	55	60	66	72	76	79
Adults	56	59	64	69	73	77	79
Memory							
Eighth grade	40	45	54	65	75	84	89
Adults	46	51	61	74	84	92	94

Columbia Graphophone Company. $1.25 per record.

References: Seashore, C. E. "The Measurement of Musical Talent," *Musical Quarterly*, 1:129–48, January, 1915.

———. *The Psychology of Musical Talent*. New York: Silver, Burdett and Company, 1919. 288 p.

———. "A Survey of Musical Talent in the Public Schools," *Uni-*

versity of Iowa Studies in Child Welfare, Vol. 1, No. 2, Iowa City: University of Iowa, 1920. 36 p.

———. "Vocational Guidance in Music," *University of Iowa Monographs*, Series 1, No. 2. Iowa City: University of Iowa, 11 p.

Kwalwasser, Jacob. *Tests and Measurements in Music*. Boston: C. C. Birchard and Company, 1927, p. 1–22, 52.

Highsmith, J. A. "Selecting Musical Talent," *Journal of Applied Psychology*, 13:486–93, October, 1929.

Test of Musical Accomplishment

Jacob Kwalwasser and G. M. Ruch (1924)

The ten subtests deal with knowledge of musical symbols and terms, time and key signatures, note and rest values, recognition of syllable names and of familiar melodies from notation, and detection of pitch and time errors in a familiar melody. The total number of responses called for is 125, some of which are to be given to multiple-answer exercises and others to single-answer ones. Each part is timed separately, the total working time being forty minutes. The basis of the test was chiefly the recommendations of the Music Supervisors' National Conference supplemented to some extent by the study of music courses in a number of cities prominent for their work in that subject. Reliability is unusually high.

$$r = .97,\ \text{P.E.}_{\text{meas.}} = 6,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .04,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .12.$$

Scores based on a total of over five thousand pupils are as follows:

Grade	*Percentile*				
	10	*25*	*50*	*75*	*90*
IX–XII	78	113	156	193	219
VIII	72	100	127	163	191
VII	54	76	106	142	168

On the whole this test must be rated as probably the best for its purpose both because of its content and makeup and its high reliability.

Bureau of Educational Research and Service. 6¢ per copy, $5.00 per 100, $40.00 per 1000.

Reference: Kwalwasser, Jacob. *Tests and Measurements in Music.* Boston: C. C. Birchard and Company, 1927, p. 65–73, 107–37.

Test of Music Information and Appreciation
Jacob Kwalwasser (1927)

This test, intended for high-school and college use, consists of nine subtests as follows:

1. Classification of artists (according to field)
2. Nationality of composers
3. Composers of famous compositions
4. Classification of composers (according to types of compositions)
5. General knowledge of composers and compositions
6. Production of tone on orchestral instruments (according to means of production)
7. Classification of orchestral instruments (according to section)
8. General knowledge of instrumentation
9. General knowledge of music structure and form

A total of 250 responses is called for, of which 150 are to true-false statements and the others to single-answer and multiple-answer exercises. The time is forty minutes. It appears that this test covers the field dealt with thoroughly enough to yield rather satisfactory and diagnostic measures of ability therein, and on the whole it must be rated as the best test available for this purpose. Tentative norms are announced as follows:

Percentile	*5*	*10*	*25*	*50*	*75*	*90*	*95*
Norm	37	52	87	113	141	165	182

These are for both high-school and college students since only neglible differences were found.

Bureau of Educational Research and Service. $5.00 per 100.

Reference: Kwalwasser, Jacob. *Tests and Measurements in Music.* Boston: C. C. Birchard and Company, 1927, p. 90–98.

Music Test
T. L. Torgerson and Ernest Fahnestock (1926)
Parts A, B

Part A, which deals with theory, contains twenty exercises that require pupils to make such responses as drawing lines around an eighth note, a half rest, and so forth, writing time signatures and syllable names, identifying natural and harmonic minor scales, placing notes, and so on. Part B, on practice or ear training, contains four subtests. For the first, twelve exercises are played by the person giving the test and pupils sing the syllables and then write them. In the second are four melodies with time signatures and measure lines left out. After hearing these played pupils are to supply them. The third subtest contains eight melodies that are played, but not just as written. The measures containing the differences are to be marked. For Subtest 4 twelve exercises are played by the tester and then written by the pupils. Twenty minutes are required for Part A, twenty-five for Part B. On the latter each response must be given in ten seconds. Part A has been criticized on the ground that shorter and better methods might have been employed to accomplish the same results, and Part B because the writing of music is not very important socially. This test was based upon the content of the most widely used texts and courses of study in the subject. A coefficient of reliability of .94 has been obtained. Although these tests are intended for high-school use, the satisfactory available norms are chiefly for the elementary school. The following medians are given, however:

Grade	*VII*	*VIII*	*IX*
Part A	17	21	—
Part B	16	18	23

Public School Publishing Company. Sample set 15¢; 75¢ per 25.

Reference: Kwalwasser, Jacob. *Tests and Measurements in Music.* Boston: C. C. Birchard and Company, 1927, p. 80–90.

Music Test Number 1

H. E. Hutchinson and L. C. Pressey (1924)

This is a test of what may be called silent reading ability in music and recognition of musical scores from well-known songs and operas. It consists of six groups of four scores each except that the last has five. With each there are the names of twice as many selections as there are scores. Pupils, of course, are to indicate from which one of the selections each score is taken. There is increase in difficulty from the first group to the last. For example, the first contains scores from *America* and *Dixie*, whereas the last contains them from the "Bridal Chorus" from *Lohengrin* and the "Soldiers' Chorus" from *Faust*. About twenty minutes' time is required, but there is no definite limit. The test may be thought of as measuring tonal imagery, which is significant in connection with successful reading of music. Failure probably means either unfamiliarity with the compositions or inability to hear mentally the sounds denoted by the musical symbols. To insure that pupils understand what is to be done, a trial group of scores and names precedes the test proper. Tentative norms for the middle of the school year based on less than a thousand cases in all are as follows:

Grade	*VII*	*VIII*	*IX*	*X*	*XI*	*XII*
Norm	6	8	10	11	11	12

Public School Publishing Company. Sample set 10¢; 50¢ per 25.

Reference: Kwalwasser, Jacob. *Tests and Measurements in Music.* Boston: C. C. Birchard and Company, 1927, p. 74–79.

II. Art

In art, as distinct from drawing, there are only three tests that seem to merit inclusion here. The first of these, as will be seen from its description, involves considerable drawing, the second somewhat less and the third none at all, but only discrimination. Thus there is no testing instrument available which attempts to measure performance in art other than along the lines of freehand drawing.

Tests in Fundamental Abilities of Visual Art
A. S. Lewerenz (1927)
Parts I, II, III

The three parts contain a total of nine subtests, as follows:

Part I, Test 1. Recognition of proportion
Test 2. Originality of line drawing
Part II, Test 3. Observation of light and shade
Test 4. Knowledge of subject-matter
Test 5. Visual memory of proportion
Part III, Test 6. Analysis in cylindrical perspective
Test 7. Parallel perspective
Test 8. Analysis in angular perspective
Test 9. Recognition of color

Test 1 contains fifteen sets of four representations of the same object or scene and requires the pupils to indicate the best of each set. Test 2 provides ten rectangles in each of which are from three to eighteen dots. Pupils are instructed to "draw some pleasing, well-proportioned shape" in each rectangle, so that the drawing includes all the dots in the space, and to label it with a single word that tells what it is. The third test presents ten drawings involving shade and shadow, and directs pupils to mark wherever they think there should be shades or shadows. Test 4 includes several matching exercises of ten items each, dealing with materials, processes, drawing terms, artists and their paintings, and so forth. For the fifth test pupils are permitted to observe a vase pictured in solid black for two minutes and are then allowed five minutes to draw its outline upon a sheet already containing lines exactly representing the top and bottom. Tests 6, 7, and 8 contain exercises requiring pupils to indicate their knowledge of perspective by marking incorrect lines in drawings. In the last test various colors are shown and pupils instructed to name the standard color which each most resembles.

Several of the tests include practice exercises. Time limits are provided for each test, the total time required being at least two hours. It is, therefore, ordinarily desirable to give the three parts at as many different sittings. The scoring of several of the tests is entirely objective, but that of others is not. For Test 2 a set of six rating sheets with which pupils' drawings are compared must be employed. In Tests 3, 6, 7, and 8 occasional uncer-

tainty is caused by the fact that pupils' marks may not be exactly placed. Test 5, which requires the outline drawing of the vase, is scored by means of a transparent key that permits slight, but very slight, opportunity for differences in opinion.

These tests are based on a survey of all available references on art tests, drawing scales, and so forth, and an analysis of the Los Angeles Visual Arts Course of Study. They are designed to measure "abilities rather than the product of abilities," or, in other words, are for prognosis rather than the measurement of achievement. Results for 1,100 pupils show a correlation of .40 with teachers' marks and of .87 between original scores and a retest a month later. For a small number of cases an average correlation of .50 with three other art tests was obtained. Median scores have been reported as follows:

	Test									
Grade	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Total*
XII	8.4	2.0	16.5	14.0	12.8	9.4	3.3	2.6	33.2	102.2
XI	7.3	2.2	16.3	14.7	10.0	8.9	2.5	3.6	32.8	98.3
X	7.2	2.1	15.3	14.2	10.0	8.6	3.8	3.6	33.5	98.3
IX	7.3	1.7	14.9	14.2	10.1	8.1	3.5	2.9	34.0	96.7
VIII	5.8	2.1	13.6	12.5	9.5	8.3	2.4	2.5	32.6	89.3
VII	4.7	1.9	13.3	11.2	9.8	7.1	2.5	1.9	33.4	85.8

Also the following ratings of art ability for junior and senior high-school pupils according to test scores are given:

	Junior High School								
	Test								
Ability rating	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Very superior	10–14	4–5	20–24	20–50	19–36	11–12	6–7	6–10	37–46
Superior	8–9	3	17–19	15–19	13–18	9–10	4–5	4–5	35–36
Average	5–7	2	12–16	11–14	8–12	7–8	2–3	2–3	30–34
Inferior	4	1	6–11	5–10	4–7	5–6	1	1	21–29
Very inferior	0–3	0	0–5	0–4	0–3	0–4	0	0	0–20

Research Service Company. $7.50 per 100; manual 35¢.

Reference: Lewerenz, A. S. "Sex Differences on Ability Tests in Art," *Journal of Educational Psychology*, 19:629–35, December, 1928.

	Senior High School								
	Test								
Ability rating	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Very superior	11–14	4–5	21–24	23–50	21–36	11–12	7	7–10	37–46
Superior	9–10	3	18–20	17–22	14–20	10	5–6	5–6	35–36
Average	7–8	2	14–17	13–16	8–13	8–9	3–4	3–4	30–34
Inferior	5–6	1	9–13	6–12	4–7	6–7	1–2	1–2	21–29
Very inferior	0–4	0	0–8	0–5	0–3	0–5	0	0	0–20

Art Ability Test

Alma J. Knauber and Luella C. Pressey

Parts A, B

Although this test is still in process of construction and apparently will be considerably modified when it appears in final form, it seems worth while to give a brief description of its present form. Part A consists of seven exercises requiring the drawing of a design from memory, the completion of drawings, the selection of the most artistic one of several objects, and so forth. Part B has ten exercises. One of these is on art vocabulary and presents forty terms, such as "blocking in," "ellipse," "hue," "low relief," and so forth, for each of which the proper one of five definitions is to be chosen. The other nine require original drawings, enlargement of a drawing, rearrangement of given material to make a design, indicating errors in drawings, and so forth. There are no time limits, but each part appears to require at least twenty-five or thirty minutes. Parts A and B are intended in their final form to measure skill alone. Others are planned to deal with art vocabulary and appreciation of beauty in everyday things. Preliminary forms of these have already been prepared, but apparently are so very tentative that they will not be described here.

The authors started with about fifty separate tests and after trying them out with several hundred university art students, arrived at those now employed. Scoring is not entirely objective, since in most cases it involves a comparison of the pupils' work with specimens in the manual, just as is done in scoring drawing and handwriting. For a small number of cases the following reliability figures are given:

$$r = .93,\ P.E._{meas.} = 2,\ \frac{P.E._{meas.}}{M} = .03,\ \frac{P.E._{meas.}}{\sigma} = .18.$$

These indicate distinctly high reliability for a test requiring no more time than this. Correlations of .75 to .80 with class work and about .60 with three other art tests are reported.

The table below gives norms, apparently based on only a few hundred cases, for various groups of individuals.

Percentile	Grade							Art specials		Elementary teachers
	VII	VIII	IX	X	XI	XII	XIII	IX	XIII	
90	52	68	68	76	88	102	83	104	135	107
50	38	52	56	64	73	82	75	82	116	88
10	28	42	47	48	52	55	53	72	93	77

The 90 percentile is called the upper critical score, and the 10 percentile the lower critical score. The authors recommend that those scoring above the upper critical score be encouraged to study art and those below the lower critical score be discouraged from studying it.

Art Judgment Test

N. C. Meier and C. E. Seashore (1929)

This test is published as a 125 page booklet. Each page contains two ink outline sketches differing in some more or less important feature. The person taking the test is to compare the two pictures in each pair, note the difference, and indicate which one he considers better or more artistic. To aid him in the comparison, the point wherein they differ is stated. The test can usually be completed within fifty minutes, but no time limit is set. The score is the number of correct responses divided by 125, or, in other words, the per cent correct. The pictures used in the test were secured by selecting works of art and altering some portion of each. A number of bases of alteration were employed, of which the following are representative:

Presence or absence of some significant feature
Appropriateness or inappropriateness of some detail
Concentration or dispersion of foreground objects

Omission or inclusion of member in series
Alteration of whole and part relations
Change in anatomical detail.

The authors first tried out the use of nine variations of each picture, later of four, but the use of a single variation coupled with a larger number of pictures appeared preferable. One hundred eighty-three paired pictures and variations arranged in chance order were tried out with more than a thousand students in elementary and high schools, normal schools, colleges, and art schools,[1] and the items which appeared most suitable were chosen for the final test.

This test is based on the belief that although other traits may be necessary or desirable to success as an artist, composition is the most indispensable and important. It follows, therefore, that one who has high ability along this line will probably possess sufficient amounts of the other necessary traits, and that one whose ability is low will be unable to succeed as an artist even if he does possess considerable amounts. It appears that in the case of this test, as with most others intended for vocational prognosis, the negative value is much greater than the positive. In other words, persons with sufficient knowledge and appreciation of art to enable them to make high scores may easily lack the manual skill or some other quality necessary for success along this line. On the other hand, there is probably little chance that anyone who has lived in an ordinary environment and is unable to make a good score upon the test can succeed as an artist.

The coefficient of reliability for the experimental form was .65. For the revised form it is almost certainly higher. Median scores are reported as follows for the group upon which the 183 items were tried out:

	Grade			*Art school*	
	VIII	*X*	*XII*	*Students*	*Faculty*
Median	66	72	76	82	87

It is suggested that scores be interpreted as follows:

[1] The art school faculty members were also included.

85 to 100—Exceptional, of rather rare capacity for sensing artistic quality, almost certain of success in an art career
77 to 84 —Superior, distinctly qualified for work in art, should profit greatly by training and be practically certain of success where a high degree of art judgment is required
70 to 76 —High average, with excellent training or superior manual skill, possibility of success in art
63 to 69 —Low average, mediocre success possible in fields not demanding great artistic judgment
53 to 62 —Poor, slight chance for artistic success
Below 52 —Zero, or no artistic ability at all

An average correlation of .55 with three other art tests has been reported for a small number of cases.

Bureau of Educational Research and Service. $1.00 per copy, $8.75 per 10, $28.50 per 35, $75.00 per 100; record sheets 2½¢ per copy, $2.00 per 100.

Reference: Meier, N. C., "A Measure of Art Talent," *Psychological Monographs*, Vol. 29, No. 2. Princeton, New Jersey: Psychological Review Company, 1928, p. 184–99.

III. Freehand Drawing

Although one of the earliest standardized measuring instruments was Thorndike's drawing scale, which will be described in this section, comparatively few other attempts to measure ability in this subject have been carried very far. Nothing that may properly be called anywhere nearly standardized has appeared to deal with knowledge about freehand drawing, as do one or two tests in the case of mechanical drawing. The only instruments available, therefore, are scales by which specimens of pupils' drawings are rated in a fashion similar to the procedure in the case of handwriting and English composition.

In the construction and use of such scales some of the same general considerations apply as have already been mentioned under the subjects just referred to. Probably the chief of these is that the samples included in a scale should deal with representations of the same object as do the specimens to be rated. Just as in composition this is not a necessity, but an advantage which results in scores of higher validity and reliability. On the whole, the few persons who have made drawing scales have heeded this require-

ment rather well. Unfortunately, however, the few available scales deal with such a small variety of subjects that they do not provide satisfactory standards for rating most of the objects drawn by children. It is probable that most children draw certain objects better than others because of interest therein, practice, or some other reason. Therefore ratings of their general drawing ability by means of specimens concerned with one or a very few assigned subjects are liable to deviate considerably from their average ability.

Scale for General Merit of Children's Drawings
E. L. Thorndike (1913)

Although this, one of the earliest standard scales, was first published in 1913, the revised form, the one now in common use, appeared in 1923. This consists of seventy specimens arranged in order at degrees of merit from zero to seventeen, inclusive. All the integral degrees except one and also a few fractional degrees are represented by specimens. All the drawings but one or two represent houses and human beings, there being at least one of each at most of the steps on the scale. Although the drawings of pupils in Grades III–VIII were employed in the derivation of the scale, the specimens at the upper degrees of merit are good enough that they may be used for rating drawings by high-school pupils. No particular directions are given by which specimens of pupils' work are to be secured. It is probable that if children's drawings are of the same subjects represented in the scale, ratings are at least as reliable as those given by any other drawing scale or by a handwriting or composition scale. The specimens given at the various steps on the scale have been rated by enough judges that the probable error of about half of them is less than .15 of the interval from one step to another. The scale is now published in pamphlet form, which is perhaps not as convenient as the original large sheet form, since only a very limited portion of it can be seen at once.

Bureau of Publications. 50¢ per copy.

References: Thorndike, E. L. "The Measurement of Achievement in Drawing." *Teachers College Record*, 14:345–83, November, 1913.

——— "A Scale for General Merit of Children's Drawings," *Teachers College Bulletin*, Series 15, No. 6. New York: Teachers College, Columbia University, 1923. 30 p.

Childs, H. G. "Measurement of the Drawing Ability of Two Thousand One Hundred and Seventy-Seven Children in Indiana City School Systems by a Supplemented Thorndike Scale," *Journal of Educational Psychology*, 6:391–408, September, 1915.

Brooks, F. D. "The Relative Accuracy of Ratings Assigned with and without Use of Drawing Scales," *School and Society*, 27:518–20, April 28, 1928.

A Measuring Scale for Free-Hand Drawing
L. W. Kline and Gertrude L. Carey (1922)
Part I—Representation (Revised)

When this appeared it was announced that other parts dealing with design and composition and color would be completed later, but these have not yet appeared. Part I contains four scales suitable for use from the kindergarten through high school, dealing respectively with a house, a rabbit, a boy running, and a brush drawing of a tree. Each scale consists of fourteen specimens, ranging from 0, or no merit at all, by intervals that are usually five, ten, or fifteen up to a specimen rated somewhere near 100, or perfection. Under each is a short paragraph pointing out the particular merits of the drawing, and also suggesting further drawings and other exercises connected therewith to the pupils. The scales are in such form that all of each may be before the eye at one time. No particular directions are given for securing samples, but there are satisfactory directions for both pupils and teachers as to how to compare drawings with the scale and assign ratings to them. If the objects drawn are the same as those used in the scales, the reliability of ratings should be fairly satisfactory. There is no reason, however, to think that these scales are superior to the Thorndike Drawing Scale except in so far as they deal with four different subjects rather than two, and as the brief paragraphs accompanying the drawings aid in using the scales. The specimens finally included in the original scales represent a selection with the aid of ninety-two judges from more than five thousand children's drawings. In making the revision, many additional judges were used, with the result that a number of new specimens were inserted and the scale values of those used previously determined more exactly.

Johns Hopkins Press, 30¢ per set of scales; 60¢ per set with account of derivation, and so forth.

References: Kline, L. W. and Carey, G. L. "A Measuring Scale for Free-Hand Drawing, Part I.—Representation," *Johns Hopkins University Studies in Education*, No. 5. Baltimore: Johns Hopkins Press, 1922. 61 p.

———— "The Kline-Carey Measuring Scale for Free-Hand Drawing, Part I.—Representation," *Johns Hopkins University Studies in Education*, No. 5a. Baltimore: Johns Hopkins Press, 1923. 10 p.

Brooks, F. D. "The Relative Accuracy of Ratings Assigned with and without Use of Drawing Scales," *School and Society*, 27:518–20, April 28, 1928.

In addition to the two instruments just described, it seems in place to mention another rather briefly, even though it is not recommended that it be used for the measurement of drawing ability. Miss Goodenough has constructed an intelligence test [2] in which the scores are based upon the drawing ability manifested by those being tested. In this test pupils are directed to draw a picture of a man upon a given sheet. Their drawings are then rated according to rather complete directions dealing with each of about fifty different points in sufficient detail to render scoring at least as objective as in the case of most rating scales. Scores are transmuted into mental ages and I.Q.'s that correlate from about .55 up to .85 with those from individual intelligence tests.

BIBLIOGRAPHY

I. Music

Baldwin, Ralph. "Efficiency in School Music Teaching and Practical Test of Same," *Music Supervisors' National Conference Proceedings*, Vol. 7, 1914, p. 43–50.

Beach, F. A. "Demonstration of the Beach Standardized Music Tests. Series I, For Achievement in Music in the Public Schools," *Music Supervisors' National Conference Proceedings*, Vol. 14, 1921, p. 186–91.

Broom, M. E. "A Note Concerning the Seashore Measures of Musical Talent," *School and Society*, 30:274–75, August 24, 1929.

Brown, A. W. "The Reliability and Validity of the Seashore Tests of Musical Talent," *Journal of Applied Psychology*, 12:408–76, October, 1928.

Kwalwasser, Jacob. "Scientific Testing in Music," *Proceedings of the National Association of Music Teachers*, Vol. 20, 1925, p. 155–64.

————. *Tests and Measurements in Music*. Boston: C. C. Birchard and Company, 1927. 146 p.

[2] This test may be secured from the World Book Company at 60¢ per 25. A complete description will be found in the following:

Goodenough, F. L. *Measurement of Intelligence by Drawings*. Yonkers, New York: World Book Company, 1926. 177 p.

Mosher, R. M. "A Study of the Group Method of Measurement of Sight-Singing," *Teachers College, Columbia University, Contributions to Education*, No. 194. New York: Bureau of Publications, Teachers College, Columbia University, 1925. 75 p.

Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction*. Yonkers, New York: World Book Company, 1927, p. 192–97.

Schoen, Max. "Common Sense in Music Testing," *Proceedings of the National Association of Music Teachers*, Vol. 20, 1925, p. 164–73.

————. "Tests of Musical Feeling and Musical Understanding," *Journal of Comparative Psychology*, 5:31–52, February, 1925.

Seashore, C. E. "The Rôle of a Consulting Supervisor in Music," *Eighteenth Yearbook of the National Society for the Study of Education*, Part II. Bloomington, Illinois: Public School Publishing Company, 1919, p. 111–23.

Streitz, Ruth. "How Are Children Selected for the Study of Music?" *High School Conference Proceedings*, 1923. Urbana: University of Illinois, 1924, p. 325–31.

Symonds, P. M. *Measurement in Secondary Education*. New York: The Macmillan Company, 1927, p. 206–13.

Trabue, M. R. "Scales for Measuring Judgment of Orchestral Music," *Journal of Educational Psychology*, 14:545–61, December, 1923.

Wright, F. A. "The Correlation Between Achievement and Capacity in Music," *Journal of Educational Research*, 17:50–56, January, 1928.

II. Art

Karwoski, T. F. and Christensen, E. O. "A Test for Art Appreciation," *Journal of Educational Psychology*, 17:187–94, March, 1926.

Reymert, Mrs. A. R. "Some Factors of Aesthetic Judgment," *Journal of Applied Psychology*, 6:34–58, 120–40; March, June, 1922.

Thorndike, E. L. "Tests of Esthetic Appreciation," *Journal of Educational Psychology*, 7:509–22, November, 1916.

III. Freehand Drawing

Cohen, Joseph. "The Measure of Drawing Ability," *Pedagogical Seminary*, 27:137, June, 1920.

Cox, G. J. "Shall We Have Intelligence Tests in Art?" *Teachers College Record*, 28:690–95, March, 1927.

Manuel, H. T. "Talent in Drawing," *School and Home Education Monograph*, No. 3. Bloomington, Illinois: Public School Publishing Company. 152 p.

Starch, Daniel. *Educational Measurements*. New York: The Macmillan Company, 1916, p. 163–70.

Symonds, P. M. *Measurement in Secondary Education*. New York: The Macmillan Company, 1927, p. 213–16.

CHAPTER XII

COMMERCIAL SUBJECTS

Introduction.—The tests to be discussed in this chapter will be grouped under six heads as follows: commercial arithmetic, bookkeeping, stenography, clerical work, business law, and general. The first three of these phases of commercial work are rather commonly, and business law somewhat less frequently, taught in high schools that offer any work in this field, but the other two are not. There seems to be a place for tests of ability to do clerical work and of a more general business nature, however, in connection with the work of the school.

In selecting the tests to be included in this chapter one encounters the difficulty that most of the relatively large number of tests in the commercial subjects have been constructed by personnel workers, employment directors, and others whose chief interest is in the business world rather than in education. Most of the tests so constructed have never been employed in any direct connection with the public schools, being used chiefly by employers, including Civil Service Commissions and other such bodies, to aid them in the selection, assignment, and promotion of employees. The examination of many of these tests reveals no fundamental differences between them and the comparatively small number more directly intended for educational purposes, so that it is not at all easy to decide which should be mentioned in this chapter and which should not. In general, however, only those have been included that seem to have a fairly close connection with actual school work.

In attempting to measure ability in such subjects as commercial arithmetic, bookkeeping, stenography, and some portions of clerical work, the test maker is not faced with the difficulties mentioned in several of the preceding chapters as existing in the case of certain school subjects. In these fields it is possible to reproduce approximately in tests the same situations and to call for the exercise

of the same abilities encountered in actual practice. The desirability of measuring such intangible outcomes as appreciation, ideals, and so forth, does not arise. Moreover, it is easier than in many other cases to compare test results with actual ability manifested in the traits measured and thus determine the validity of the tests.

Although the majority of the tests described in this chapter deal with ability or achievement in the various commercial subjects, a few of them are prognostic in their nature. These attempt to measure the capacity of individuals to become efficient stenographers, clerks, and so forth, regardless of how much or little they actually know about the details of the work itself. Such tests naturally have their chief value in connection with vocational guidance, but because they belong in the commercial field have been dealt with here rather than in the portion of a later chapter devoted to general vocational guidance.

I. Commercial Arithmetic

It is rather surprising that the number of tests in commercial arithmetic is so small. There is only one that seems to the writer to deserve recommendation for school use. Portions of some tests mentioned in other sections of this chapter contain a few problems in commercial arithmetic, but are scarcely long enough to be used as complete tests therein. Also a number of the standardized arithmetic tests intended for use in the upper elementary grades, especially those dealing with written problems, contain exercises that might properly be classed as commercial arithmetic and in many cases difficult enough that they might well be used for testing in the secondary school. The reader who is interested in these, however, is referred to one of the many books devoted chiefly to elementary-school tests.

Scale of Problems in Commercial Arithmetic

L. B. Kinney (1926)

Test A, Parts 1, 2; Tests B, C, Forms 1, 2, of each

Test A is to be used at the end of ten weeks, B at the end of the first semester, and C at the end of the second. Part 1 of Test A includes four subtests, one dealing with each of the four fundamental operations. The first has forty addition examples, each

consisting of eight two-place numbers; the second sixty subtraction examples, all of the numbers concerned being three-place; the third sixty multiplications of two- or three-place numbers by two-place numbers; and the last forty divisions of three- or four-place numbers by 10, 25, 50, 75, 100, or 1,000. The second part of Test A, on aliquot parts, consists of twelve exercises, each of which contains five similar problems dealing with cost of goods, amount of interest, and similar situations. Each form of Tests B and C includes ten written problems of the type commonly dealt with in high-school commercial arithmetic. Five minutes are allowed for each subtest of Part 1 of Test A, ten for Part 2 and "what time they need" for B and C. It is stated that the attempt was made to include problems "containing a minimum of linguistic difficulty, a maximum of concreteness of situation." The problems were collected from both business men and teachers.

These tests are diagnostic to a certain degree. They do not diagnose the elemental abilities that enter into total commercial arithmetic ability, but they do break the latter up into certain chief parts and provide enough examples under each that they have considerable diagnostic value. An unusual feature intended to aid in their use for this purpose is that on the last sheet of each test provision is made for pupils to record and tabulate the examples or problems missed, and in the case of Tests B and C to record the reason why each was missed. The coefficients of reliability of the parts of Test A range from .77 to .89, but those of Tests B and C are not quite .70.

The scoring directions do not state whether the omission of dollar marks and per cent signs should be penalized, or if both decimals and common fractions are satisfactory. Since many of the answers contain these elements, it is probable that different scorers will not agree on the scores they assign. The suggestion is made that scores on Test B and C, which are in terms of numbers of problems correct, be turned into point scores according to a table given which allows for the difference in difficulty of the various problems. T-scores are also available for all three tests.

The table for converting raw into corrected scores for Tests B and C indicates that in both cases the second forms are roughly 10 per cent more difficult than the first ones. Only tentative norms have been announced so far. These are as follows:

Test A	93.1
Part 1	80.6
Addition	11.5
Subtraction	42.3
Multiplication	12.5
Division	14.3
Part 2	12.5
Test B	5.4
Test C	5.5

From these it appears that the subtraction subtest does not contain enough examples to measure satisfactorily the ability of all pupils since the norm or median is at about 70 per cent of the number of examples included, and able pupils can complete the entire number of examples in less than the time allowed.

Public School Publishing Company. Sample set 25¢; $1.00 per 25.

Reference: Kinney, L. B. "Measurement of Results of Teaching in Commercial Arithmetic," *University of Iowa Monographs in Education*, Series I, No. 7. Iowa City: University of Iowa, 1926, p. 96–112.

II. Bookkeeping

Although the number of tests in bookkeeping is somewhat greater than in the case of commercial arithmetic, it is much smaller than is true for stenography and clerical work. In addition to the very few general ones, however, many others that have not at all approached standardization are published in connection with particular textbooks and manuals in bookkeeping, and may well be employed by teachers.

Bookkeeping Tests

P. A. Carlson (1925)

Series A, Tests 1, 2, 3, 4, 5, 6, 7, 8, 9; Series B, Tests 1, 2

Although these tests are especially for use with the same author's *Twentieth Century Bookkeeping*, their content is such that they may be employed with other texts as well. They are intended to be given at different times, Test 1 of Series A covering approximately the first six weeks' work, and each of the others follows in order. The two Series B tests cover about twelve and twenty-four weeks' work, respectively. Each test consists of from three to

six sections dealing with different phases of the work, both theory and practice, and containing exercises of various types. For example, Series A, Test 1, includes four sections, the first of which deals with journalizing and the second with posting, the third contains true-false statements on the theory of bookkeeping, and the fourth completion statements on the same topic. To give a second example, Test No. 9 of Series A deals with adjusting entries, classification of accounts, the working sheet, numerical classification of accounts and closing and posting entries. The number of items on the various tests ranges from about one hundred up to more than twice that number. No definite time limits are given, but instead three possible plans of administering the test are suggested. (1) The teacher may give pupils time to finish, but note the time consumed by each. This is stated to be most frequently used. (2) The teacher may allow all pupils to finish without taking any record of time and consider only accuracy in scoring. (3) The teacher may instruct the class to write only one page at a time and as soon as the first pupil has finished each page, direct the whole class to begin the next page. Of these three methods the first is probably to be preferred. An ordinary high-school period should be sufficient for most pupils to complete any one of the tests.

In a few cases the answers given as correct are not in agreement with all texts and instructors. Hence perfect objectivity of scoring is not secured. Norms of accuracy based upon more than twenty thousand pupils in about one thousand schools for Tests 1 and 2 of Series A, and on about half as many pupils and schools for Tests 3 and 4 of the same series have been reported as follows:

Test	*Percentile*						
	5	*10*	*20*	*50*	*80*	*90*	*95*
1	66	73	81	92	101	105	107
2	55	62	70	83	91	94	96
3	64	72	81	97	111	116	119
4	43	48	54	65	77	82	87

Similar norms are also given for each section of each test and time results for Tests 3 and 4.

South-Western Publishing Company. No charge except postage.

Reference: Carlson, P. A. "Some Problems of Measurement in Bookkeeping," *University of Iowa Monographs in Education*, Series I, No. 7. Iowa City: University of Iowa, 1926, p. 152–60.

Bookkeeping Tests[1]

(1926)

Series I, II, III, IV, V, VI, VII, VIII, IX; 6 tests in each

These tests were prepared primarily for use with Jackson, Sanders and Sproul's *Bookkeeping and Business Knowledge*. The six tests in each series are true-false, completion, association, selection, enumeration, and reasoning. Some of those in each series are highly objective, whereas others are rather of the traditional type. Apparently each series covers certain chapters, although this is not stated on all the tests. The first five series deal with the first year's work, and the last four with that of the second year. Time limits are not stated, but apparently it is intended that teachers giving the tests use as much time as they think best. The true-false, completion, and association tests do not appear to require more than ten minutes each, whereas for the others probably more time should be allowed.

Ginn and Company. No charge except postage to those employing Jackson, Sanders and Sproul's text. To others $1.20 per 15 of each series.

Rational Objective Tests in Bookkeeping and Accounting

Gregg Publishing Company (1928)

Series A, Tests 1, 2, 3, 4, 5, 6, 7, 8, 9, 10; Sections 1, 2 of each

Although not primarily intended for high-school use, these tests appear to be suitable therefor. They deal with the following topics in order: asset accounts, deferred entries, proprietorship, receivables, liabilities, trial balance and clerical errors, the journals, sales and purchase discounts, interest and discount, business practice and procedure. Section 1 of each test contains twenty-five statements to be marked true or false, and Section 2 the same number of completion statements, most of which have to do with

[1] No author is named for these tests.

definition or explanation of terms. No set time limits are provided, but each pupil's time is to be recorded. The tests are intended to cover approximately equal portions of the subject matter commonly taught in first-year bookkeeping and, therefore, may be given at about the end of each month of school. On the whole, they appear to have considerable merit, although for as comprehensive a series as they compose, it would probably be better to employ additional types of exercises including some actual examples as they occur in practice.

Gregg Publishing Company. Complete set 20¢; 2¢ per copy; key 10¢.

Bookkeeping Test

F. H. Elwell and J. G. Fowlkes (1928)

Tests 1, 2; Forms A, B of each

Test 1 is intended to be used at the end of the first semester and Test 2 at the end of the second. Each contains nine parts. The first three of these deal with general theory, the fourth with journalizing, the fifth with bookkeeping, the sixth with adjusting entries and closing the ledger, and the last three with statements. They include a total of ninety-five test elements in alternative, multiple-answer, and completion form. The time is fifty minutes.

The content of the test was determined by a study of first-year bookkeeping courses throughout the country. The present forms have resulted from tryouts of three experimental editions and criticism by both teachers of bookkeeping and educators. Reliability data are as follows:

Test 1:

$$r = .82,\ \text{P.E.}_{\text{meas.}} = 4,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .05,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .29.$$

Test 2:

$$r = .87,\ \text{P.E.}_{\text{meas.}} = 4,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .06,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .24.$$

The announced norms are based on only a few hundred pupils for each test. They are as follows:

Test	Percentile						
	5	10	25	50	75	90	95
1	49	55	67	79	87	94	97
2	40	47	56	70	81	90	93

These are averages for the two forms. In the case of Test 1 there is very little if any difference in difficulty, but in that of Test 2 scores on Form A are usually about two points above the figures given and those on Form B about two points below.

Because of the length of the test, the basis of selecting items, the fairly high reliability, and other desirable qualities, the writer believes that this is the best test available for general school use in this subject.

World Book Company. Specimen set 30¢; $1.30 per 25.

Bookkeeping Achievement Tests

C. E. Bowman (1929)

Tests 1, 2, 3, 4, 5, 6

This series of tests is based upon Bowman and Percy's *Principles of Bookkeeping and Business, Elementary Course*. Each covers several of the chapters of the text just named in order. The number of elements in each is about fifty or sixty, grouped in two parts, of which the first deals with the principles and the second with their application. Directions provide for allowing all the time needed and noting that taken by each pupil. Apparently no one of the tests requires as much as one ordinary class period.

American Book Company. 96¢ per 15 copies of complete series.

Bookkeeping Tests

F. H. Elwell and J. V. Toner (1929)

Sets I, II, III, IV, V, VI, VII, VIII; 4 tests in each

The four tests in each set are true-false, matching, multiple-answer, and completion. Each group covers from six to ten chapters in Elwell and Toner's *Bookkeeping and Accounting*. All of each test is contained on a single side of one sheet and should be completed by most pupils within from five to ten minutes. These

tests are so closely adapted to the text upon which they are based that it is doubtful if they should be employed in connection with other texts except perhaps at the completion of the course.

Ginn and Company. No charge except postage to those employing the Elwell and Toner system. 80¢ per 15 of all four tests.

III. Stenography

As employed here, the term "stenography" is intended to include both typing and shorthand. Of the numerous tests constructed for use in the business world and the few for school, only a small number will be described in this section. The chief reason for this is that there is very much similarity among the different tests so that it is scarcely worth while to include more than a few to indicate what most of them are like. It is somewhat easier to test typing than shorthand ability, especially when the attempt is made to reproduce actual working conditions. However, some of the tests that deal with shorthand appear to possess as high validity as those in almost any of the school subjects.

Examination in Typing

L. L. Thurstone (1920)

Form A

This, originally published under the title "Proficiency Test for Typists," was one of the earliest standardized stenographic tests. It consists of three parts. In the first is a reproduction of a two-hundred word typed letter, revised and corrected in longhand, which is to be copied in correct form. The second part consists of ten items each containing the name of a state, an amount of money, a statement as to means of shipping and a date. These are to be tabulated and typed. In the third part are forty-eight words of which twelve misspelled ones are to be crossed out. No time limit is set, but the time consumed by an individual taking the test is recorded. Apparently few persons should require more than thirty minutes and many less than that. Since this test covers only a limited portion of a stenographer's work, it would not appear to possess high validity as a general measure of stenographic abil-

ity, but for the types of work covered is probably rather satisfactory.

Two scores are provided for, one for errors and the other for speed, and also a combined score. They are interpreted as follows:

Error score	*Total time*	*Combined score*	*Rating*	*Interpretation*
0–10	14 or less	30 or less	*A*	Very inferior
11–20	15–24	31–50	*B*	Superior
21–30	25–29	51–70	*C*	Average
31–50	30–34	71–80	*D*	Inferior
51 or more	35 or more	81 or more	*E*	Very superior

World Book Company. Specimen set 15¢; $1.50 per 25.

Stenographic Proficiency Tests

E. G. Blackstone (1923)

Typewriting; Forms A, B, C, D, E

Each form of the test consist of a letter requiring slightly more than one thousand strokes to type. These letters were very carefully constructed so as to be similar in seven points which might affect their difficulty. Those taking the test first prepare their machines according to standard directions and practice five minutes on other material, then have three minutes in which to make a copy of the letter. The score is determined by dividing ten times the number of strokes per minute by the number of errors plus ten, thus being a combined measure of speed and accuracy.

Within the limited field covered, that of copying a business letter, this test appears to yield rather satisfactory measures. Since work of this nature is frequently an unimportant phase of what typists do, it does not appear that the scores obtained are necessarily valid measures of general typing ability. Correlations based on about one thousand cases indicate that there is practically no correlation between scores on this test and chronological or mental age, intelligence quotient, or eighth-grade standing.

Norms based upon several thousand cases have been announced as follows:

Months studied	*5*	*10*	*15*	*20*	*25*	*30*
Score	88	148	178	200	220	236

World Book Company. Specimen set. 15¢; $1.00 per 25.

References: Blackstone, E. G. "Measurement of Progress in Typewriting," *Detroit Journal of Education*, 1:35–41, May, 1921.
Davis, H. H. "Measurement in Commercial Education in the St. Louis Schools," *University of Iowa Monographs in Education*, Series I, No. 7. Iowa City: University of Iowa, 1926, p. 42–52.

A Series of Tests in Gregg Shorthand

Elmer Hoke (1921)

Tests A–1; B–1, 2; C–1, 2, 3, 4, 5, 6, 7, 8, 9, 10;
Measuring Scales for Gregg Shorthand Penmanship and for Knowledge of Gregg Shorthand

Although this elaborate series of tests was made to be used in connection with Gregg shorthand, some of them may well be employed to measure ability in any system of shorthand. Test A, on reading ability, consists of two letters, having a total length of five hundred words, written in shorthand with every tenth word omitted. Each omission is to be supplied by indicating the correct one of two printed words. The incorrect words employed were selected from one column of the Ayres Spelling Scale[2] so as to make their choice purely random and to avoid requiring close discrimination. The B tests are on speed of writing and consist of about four hundred words presented in "tri-interlinear" form. That is, below each printed line is the correct shorthand therefor and below that a blank space on which the individual being tested is to copy the shorthand. Each of the C or vocabulary tests contains one hundred words and fifty phrases, for which the correct shorthand symbols are to be given. The words were taken from the Ayres Spelling Scale and the phrases from a study of over forty thousand phrases contained in almost half a million words of the kind of material stenographers are commonly required to take.

[2] Ayres, L. P. *A Measuring Scale for Ability in Spelling*. New York: Division of Education, Russell Sage Foundation, 1915. 59 p.

The thousand words make up 92 per cent of the words in English composition and the five hundred phrases 78 per cent of the phrases. The shorthand penmanship scale consists of sixteen specimens, the first of which is rated zero and the others by fives from twenty-five to ninety-five, inclusive. Starting with 1,155 specimens of pupils' work, the number was gradually reduced following ratings by various groups of judges, several hundred in all, until the samples actually used in the scale were selected. Provision is made for a standardized procedure in securing samples to be rated by the scale. The scale for knowledge of the system is similar in form to the Ayres Spelling Scale. It consists of the words and phrases that compose the ten vocabulary scales, arranged in twenty-one columns, those in each being of practically the same difficulty. The per cents of correct responses to be expected from pupils who have studied shorthand for a year and two or three months are given at the heads of the columns.

The directions set time limits of three and two minutes, respectively, for the reading ability and speed-of-writing tests, and allow as much time as is needed for each vocabulary test. Also they emphasize both speed and accuracy. Scores on Tests A and B are entirely objective, but those on the vocabulary tests and also the ratings according to the two scales are not, as indeed they cannot be. Apparently, however, as close an approach to objectivity is made as in the case of the best penmanship and other similar scales. Hoke reports coefficients of reliability of .80 for Test A and .85 for B, also correlations of from .50 to .75 with scores determined otherwise.

Norms have been announced as follows, based on only a few hundred pupils for Tests A and B, but on about two thousand for Vocabulary and Quality of Writing.

Semesters	*Test A*	*Test B*		*Test C* *	*Quality of writing scale* †
		1	*2*		
IV	76	70	79	118	82
III	66	70	82	122	—
II	64	56	67	112	77

* The scores given are the average scores on the ten vocabulary tests.

† The first score in this column is for pupils who have had 29 weeks of instruction, the second for those who have had 66 weeks.

From these scores it appears that B–1 is more difficult than B–2, but most of the difference is due to the fact that the B–2 scores are for pupils who had already taken B–1. The true difference, according to Hoke, is only about two points. The scores on the various vocabulary tests do not differ a great deal. Test C–1 appears to be several points harder and Tests C–4, C–5, C–7 and C–8 three or four easier than the others.

Gregg Publishing Company. Tests A, B, and C—specimen set, 30¢; 25¢ per 25. Shorthand Penmanship Scale, 25¢. Knowledge of Shorthand Scale, 10¢.

Reference: Hoke, E. R. "The Measurement of Achievement in Shorthand," *Johns Hopkins University Studies in Education*, No. 6. Baltimore, Maryland: Johns Hopkins Press, 1922. 118 p.

Prognostic Test of Stenographic Ability

E. R. Hoke.

This test consists of several subtests intended to measure the following abilities, supposedly contributory to stenographic ability: motor reaction, speed and quality of writing, speed of reading, memory, spelling, and use of symbols. Motor reaction is tested by requiring the making of marks in a number of spaces provided therefor. Speed and quality of writing are determined by having a selection copied as often as possible within the allotted time. The next test requires the reading of a selection that has every tenth word in parentheses, accompanied by an extraneous word. Those being tested are to mark the word of each pair that makes sense. The memory test requires the reproduction of twenty-five word sentences from memory. In the spelling test are sixty words each spelled twice, once correctly and once incorrectly. Finally a letter symbol is given for each of the single digits followed by a page containing two hundred digits in irregular order, for which the corresponding letters are to be filled in.

One study yielded correlations ranging from .36 to .76 with college marks in shorthand. In the same study it was found that the Terman Group Test of Mental Ability[a] yielded practically the same correlations with shorthand marks as the Hoke test. A

[a] This test is described on page 405.

similar study of high-school marks yielded a correlation of .82 with Prognostic Test scores. The scores given below are for more than twelve hundred beginning shorthand pupils in the high schools of twenty-six cities.

Percentile	*5*	*10*	*25*	*50*	*75*	*90*	*95*
Score	291	311	349	397	440	485	506

Gregg Publishing Company. 2¢ per copy.

Reference: Wood, E. H. "An Experiment with Predictive Tests in Stenography," *Journal of Commercial Education*, 57: 298–99, 318, December, 1928; 58:14–15, 24, January, 1929.

Diagnostic Shorthand Tests
Ethel A. Rollinson (1924)
Lessons I, II, III, IV

This series is intended for use with the *Gregg Shorthand Manual*. Each test consists of four parts which measure knowledge of principles, penmanship ability, facility of writing, and reading ability. The first consists of fifty words or combinations of two words, for each of which the shorthand form is to be written. The words used are, according to Thorndike's *Teachers' Word Book*,[4] the most common ones that illustrate the principles presented in each of the first four chapters of the Manual. The second part of each test places before the pupils a sentence written in longhand and, on the blackboard, the correct shorthand for it. After practicing the shorthand for one minute, the pupils spend two minutes writing it in the folder. The test on facility of writing presents a short paragraph in longhand below each line of which is the correct shorthand therefor, and below that a blank line. The pupils are to copy the shorthand. The reading ability test consists of a paragraph of shorthand material with occasional omissions of single words. Four words in longhand are given for each omission and the correct one of the four is to be underlined. Each of the tests covers practically all points in one of the first four chapters of Gregg's Manual.

[4] Thorndike, E. L. *The Teacher's Word Book*. New York: Columbia University, 1921. 134 p.

The directions to pupils and also the instructions for those giving the test are not entirely clear. On the whole, however, their deficiencies are not very serious and probably do not result in lowering objectivity very much. The total actual working time is fifteen minutes and the time limits set upon each of the four tests are short enough that for most pupils speed is a factor in determining the scores earned. The score on penmanship ability does not, however, depend upon the amount written. Although the correct answers for the first and last parts of each test and standards for rating the second and third parts are provided, scoring is perfectly objective only in the last part.

The manual which accompanies the tests states, without giving the exact figures, that knowledge of principles correlates highly with penmanship ability and also with reading ability, that its correlation with facility of writing and that between penmanship ability and reading ability are only fair, and that the other intercorrelations of the various subtests are low. It has also been found that the scores on knowledge of principles and on reading ability tend to have a rather high correlation with final marks in stenography. The correlation of penmanship ability with final marks is only medium and that of facility of writing with marks rather low. A correlation of .89 between total scores and final marks is reported. The manual contains helpful suggestions for dealing with particular classes of pupils and for the general improvement of instruction. It also gives a list of from six to nine specific aims for the tests on each of the four abilities included.

Medians for over twelve hundred pupils in twenty-six cities during the second semester are as follows:

Abilities	*Test*			
	1	*2*	*3*	*4*
Knowledge of principles	28	24	33	33
Penmanship ability	5.8	6.0	5.5	6.6
Facility of writing	29	31	33	37
Reading ability	6.2	5.0	6.5	8.2
Total	69	66	78	85

Gregg Publishing Company. Sample set 25¢; 4¢ per copy.

Objective Tests in Typewriting
Esta R. Stuart (1929)
Series A; Tests 1, 2, 3, 4, 5, 6, 7

Test 1 is intended to be given during the fifteenth week of instruction, Test 2 between the seventeenth and twentieth weeks, for the purpose of testing mental control. Tests 3 to 7 are designed to measure the accuracy and rapidity of the automatic manipulative responses of the student. The first two, which contain fifty multiple-answer and one hundred true-false elements, respectively, deal chiefly with knowledge about typewriters, and of how to use them. The other five provide exercises in actual typewriting performance. Those taking the test are allowed as much time as is needed to complete them, but the time used is recorded and apparently considered as a time score.

Gregg Publishing Company. Sample set 25¢; 2¢ per copy.

IV. Clerical Work

As has already been suggested, clerical tests are not usually tests of anything taught as such in high-school courses, but rather measure capacity and probable ability in clerical work. Some of them attempt to reproduce as exactly as possible some of the situations that clerks frequently encounter, whereas others try to measure certain traits or capacities that seem to underlie the ability to perform the activities commonly required of clerks. Those of the former type are ordinarily most appropriate for the testing of individuals who have already had clerical experience, whereas those of the latter are better suited for beginners or applicants who have not worked in this field.

There is one outstanding organization working in this field which should be mentioned although its work has been in connection with industry and business rather than the schools. This is the Bureau of Public Personnel Administration. For a number of years this bureau has rather regularly been producing at least one new test per month and many of these are concerned with different varieties of clerical work. Among the phases of such work covered are the running of an addressograph, alphabetical

filing, accounting, the duties of junior and senior clerks, and so forth.[5]

Examination in Clerical Work
L. L. Thurstone (1919)
Form A

This was originally published under the title "Clerical Examination," and as such was among the first standardized clerical tests. It consists of eight parts as follows:

1. Checking addition and subtraction errors in 120 easy examples
2. Underscoring misspelled words in a fairly difficult passage of 800 words
3. Cancelling four letters each time they occur in 300
4. Code learning—substitution for thirty-six items of four symbols each
5. Classifying and alphabetizing thirty-seven proper names
6. Classifying twenty-five insurance policies according to kind, date, and amount
7. Solving twelve arithmetic examples
8. Matching ten English with the same number of Arabian proverbs

No definite time is provided, but that required by each individual tested is noted. It usually ranges from forty to eighty minutes. Speed scores correlate .42, accuracy scores .50, and combined scores .61 with ratings of clerical workers on the quality of work actually done. Separate error or accuracy and speed ratings are given, also a combined rating. Scores are interpreted as follows:

Error score *	*Total time in minutes*	*Error score plus total time*	*Rating*	*Interpretation of rating*
0–12	49 or less	71 or less	*A*	Very superior
13–22	50–59	72–87	*B*	Superior
23–42	60–71	88–111	*C*	Average
43–62	72–79	112–131	*D*	Inferior
63 or more	80 or more	132 or more	*E*	Very inferior

* The error score is not simply the number of errors but is a weighted score.

[5] References to these tests will be found in the bibliography at the end of the chapter. Many of the tests produced by this bureau are being used in considerable numbers by employers.

World Book Company. Specimen set 15¢ $1.50 per 25.

Reference: Thurstone, L. L. "A Standardized Test for Office Clerks," *Journal of Applied Psychology*, 3:248–51, September, 1919.

Clerical Aptitude Tests, Junior Grade

L. J. O'Rourke

Clerical Problems, Reasoning Test; Form A of each

The clerical problems test contains six subtests, of which the first and last deal with alphabetical filing; the second and fifth with arithmetic; the third requires the checking of names in a list according to certain stated qualifications; the fourth, recognition of errors in copying. The reasoning test consists of forty-five exercises dealing with business information, ability to solve problems in arithmetic, spelling ability, and so forth, many if not most of which appear to test stock of information rather than reasoning ability. The working time on each test is twenty minutes. The test on clerical problems resembles a number of other tests for clerical workers and a comparison of it with them reveals no reason why it should be particularly more or less valid and reliable.

Educational and Personnel Publishing Company. $3.00 per 100.

V. Business Law

In general only those high schools that have rather strong commercial departments offer courses in business law. However, the number of these is sufficient that there should be a supply of tests available for use. In so far as the writer knows, however, only one series of such tests appropriate for use in the secondary school has appeared. Although these tests are not standardized, but are to a considerable extent of the nature of exercises and practice material, they will be described below.

Case Problems and Tests in Business Law

F. K. Beutel and C. G. Rediker (1929)

Tests 1–54

These tests cover the following ten topics: relation of law to

business, contracts, sales, bailments and carriers, insurance, loans-credit-guarantee, negotiable instruments, agency, business organization, and real and personal property. There are five tests on each topic except negotiable instruments, which has only four, and also five general review tests. Each set of five includes one on case problems, which approaches the traditional type of examination, one true-false, one selection, one matching, and one business-judgment test, which likewise is somewhat traditional in its form. Each test contains from eight to thirty exercises or items. Time limits are not given, but presumably from ten to twenty minutes are sufficient. The series is stated to cover the important points in the usual course in elementary business law and appears to do so rather thoroughly.

Ginn and Company. 52¢ per set.

VI. Commercial Geography

Very little has been done by way of attempting to measure achievement in commercial or other high-school geography. A number of the geography tests or series of tests intended for use in the elementary school contain a few elements, an entire subtest or, in the case of a series of tests, a whole test devoted to this phase of the subject. However, these are all primarily intended for use in the elementary school and, therefore, will not be named here. Just one series of tests will be described in this section.

Industrial and Commercial Geography Tests

J. W. Morris (1930)

Tests 1, 2, 3

Each test is intended to cover one six-weeks period and thus the three together cover the work of a one-semester course in this subject. There are seventy or eighty elements in completion, multiple-answer, and true-false form in each test. Test 1 is devoted to United States and 2 to the remainder of the world. The time allowed is thirty-five minutes.

Harlow Publishing Company. Sample set 10¢; 75¢ per 25, $2.50 per 100.

VII. General

There have been quite a number of so-called tests of business aptitude or some other similar ability constructed, but in so far as the writer knows, none of them have received any use in secondary schools. Many of the earlier ones are now no longer available since it has been found that they were apparently no more predictive of business success than an intelligence or perhaps some other test. As an example of what some of them are like, two of the best of the recent ones will be described in this section.

Business Aptitude Test

F. A. Moss, K. T. Omwake, and Thelma Hunt (1926)

This has six parts. Test 1, observation and judgment, contains forty true-false items having to do with general business information concerning well-known automobiles, typewriters, important commercial centers, taxation, law, and so forth. Test 2, memory for names and faces, contains the photographs of twenty-five individuals with the names of twelve of them. Photographs of these twelve are shown before Test 1 is taken and when Test 2 is reached these twelve are to be selected and connected with the proper names. Test 3, comprehension, calls for matching ten proverbs with others having the same meanings. The fourth, on reasoning, requires the solution of six written arithmetic problems. The fifth, meaning of words, consists of thirty pairs of words to be marked as meaning the same or the opposite. In Test 6, following directions, is a table containing fifteen titles of employees with five items of information concerning each. Following these are three blank columns in which symbols are to be placed according to six different directions accompanying the table. The directions allow five minutes in which to study the sheet of names and faces and thirty minutes working time upon the test. The subtests may be timed separately or not, at the option of the person giving the test. This, of course, tends to make results somewhat less comparable than if all testers followed exactly the same procedure.

$$r = .88,\ \text{P.E.}_{\text{meas.}} = 6,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .08,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .24.$$

Norms for high-school seniors entering business college based on two or three thousand cases, and for students qualified for business positions are as follows:

	Percentile				
	10	*25*	*50*	*75*	*90*
Qualified students	44	58	75	96	116
High-school seniors	41	52	66	80	94

Center for Psychological Service. 7¢ per copy, $5.00 per 100.

General Business Training Tests

J. R. Smith (1929)

Tests 1, 2, 3, 4, 5, 6, 7, 8

The tests of this series are prepared especially to cover Crabbe and Slinker's *General Business Training*. Each consists of four sections in true-false, completion, multiple-choice, and matching forms, and contains a total of sixty elements. The total time for each test is about forty minutes, but no definite limit is given. Norms for about three thousand pupils from one hundred different schools are reported as follows:

Test	Percentile						
	5	*10*	*25*	*50*	*75*	*90*	*95*
1	32	35	41	47	52	56	58
2	31	36	42	48	52	55	56
3	31	35	41	46	51	54	56
4	40	42	46	49	52	54	56
5	32	34	39	44	49	52	53
6	36	40	44	48	52	54	56

South-Western Publishing Company. No charge to those using Crabbe and Slinker's text.

BIBLIOGRAPHY

I. Commercial Arithmetic

Cody, Sherwin. *Commercial Tests and How to Use Them.* Yonkers, New York: World Book Company, 1920, p. 97–106.

II. Bookkeeping

Carlson, P. A. "Some Problems of Measurement in Bookkeeping," *University of Iowa Monographs in Education*, Series 1, No. 7. Iowa City: University of Iowa, 1926, p. 152–60.

Tanz, Louis. "New Type of Tests in Bookkeeping," *Bulletin of High Points*, 6:22–27, December, 1924.

III. Stenography

Baugh, R. D. "Objective Measurement Tests in Commercial Subjects," *The High School*, 4:131–34, May, 1927.

Bills, M. A. "Methods for the Selection of Comptometer Operators and Stenographers," *Journal of Applied Psychology*, 5:275–83, September, 1921.

———. "A Test for Use in the Selection of Stenographers," *Journal of Applied Psychology*, 5:373–77, December, 1921.

Book, W. F. "How Progress in Learning to Typewrite Should be Measured and Why," *University of Iowa Monographs in Education*, Series 1, No. 7. Iowa City: University of Iowa, 1926, p. 62–76.

Burt, Cyril. "Tests for Clerical Occupations," *Journal of the National Institute of Industrial Psychology*, 23–27:79–81, 1922.

Chapman, J. C. and Chapman, Daisy. *Trade Tests*. New York: Henry Holt and Company, 1921. 435 p.

Cody, Sherwin. *Commercial Tests and How to Use Them*. Yonkers, New York: World Book Company, 1920, p. 152–66.

Cook, W. A. "An Effort to Measure Typing Efficiency," *Journal of Educational Research*, 11:49–59, January, 1925.

Dush, W. M. "The Building and Use of Achievement Tests in Gregg Shorthand," *Journal of Commercial Education*, 57:114–15, April, 1928.

Easterbrook, Mabel and Navra, M. A. "Factors Predetermining Success in Typewriting," (*St. Louis*) *Public School Messenger*, 21:8–20, May, 1924.

Freyd, Max. "Selection of Typists and Stenographers: Information on Available Tests," *Journal of Personnel Research*, 6:490–510, April, 1927.

———. "The Selection of Typists—Information on Available Tests," *Journal of Commercial Education*, 56:133–37, 167–69; May, June, 1927.

———. "Tests for Stenographers," *Journal of Commercial Education*, 56:200–3, September, 1927.

———. "Tests for Typists and Stenographers," *Journal of Commercial Education*, 56:236–37, October, 1927.

Gronert, M. L. "A Prognostic Test in Typewriting," *Journal of Educational Psychology*, 16:182–85, March, 1925.

Harned, W. E. "Tests and Measurements in Relation to Typewriting," *Fifteenth Annual Schoolmen's Week Proceedings*. Philadelphia: University of Pennsylvania, 1928, p. 550–55.

Hoke, E. R. "Educational Measurements and the Teaching of Shorthand,"

Fifteenth Annual Schoolmen's Week Proceedings. Philadelphia: University of Pennsylvania, 1928, p. 555–60.

Jessup, E. M. "The Application of Prognostic and Achievement Tests to Shorthand," *Journal of Commercial Education,* 57:173–74, June, 1928.

Nies, F. E. "Unit Measurement of Shorthand," *University of Iowa Monographs in Education,* Series 1, No. 7. Iowa City: University of Iowa, 1926, p. 122–29.

Ohmann, O. A. "The Measurement of Capacity for Skill in Stenography," *Psychological Monograph,* Vol. 36, No. 2. Princeton, New Jersey: Psychological Review Company, 1926, p. 54–70. Also in: *University of Iowa Studies in Psychology,* Vol. 10. Iowa City: University of Iowa, 1926.

————. "The Possibility of Prognosis in Stenography," *University of Iowa Monographs in Education,* Series 1, No. 7. Iowa City: University of Iowa, 1926, p. 36–41.

Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction.* Yonkers, New York: World Book Company, 1927, p. 191–92.

Shellow, S. M. "An Intelligence Test for Stenographers," *Journal of Personnel Research,* 5:306–8, December, 1926.

Stedman, M. B. "A Study of the Possibility of Prognosis of School Success in Typewriting," *Journal of Applied Psychology,* 13:505–15, October, 1929.

Symonds, P. M. *Measurement in Secondary Education.* New York: The Macmillan Company, 1927, p. 221–28.

Tuttle, W. W. "The Determination of Ability for Learning Typewriting," *Journal of Educational Psychology,* 14:177–81, March, 1923.

Vavra, M. A. "Success in Typewriting," *Journal of Educational Psychology,* 16:487–92, October, 1925.

Wood, E. H. "An Experiment with Predictive Tests in Stenography," *Journal of Commercial Education,* 57:298–99, 318, December, 1928; 58:14–15, 24, January, 1929.

"Preliminary Work on Tests for Stenographer," *Public Personnel Studies,* 6:46–55, February, 1928.

IV. Clerical Work

Cody, Sherwin. *Commercial Tests and How to Use Them.* Yonkers, New York: World Book Company, 1920, p. 57–69, 167–74.

Moss, F. A. and Telford, Fred. "Suggested Tests for Senior Clerk," *Public Personnel Studies,* 2:195–213, September, 1924.

Ruggles, A. M. "A Diagnostic Test of Aptitude for Clerical Office Work, Based on an Analysis of Clerical Operations," *Teachers College, Columbia University, Contributions to Education,* No. 148. New York: Bureau of Publications, Teachers College, Columbia University, 1924. 93 p.

Symonds, P. M. *Measurement in Secondary Education.* New York: The Macmillan Company, 1927, p. 376–83.

Telford, Fred and Moss, F. A. "Suggested Tests for Supervising Clerk," *Public Personnel Studies,* 2:288–97, December, 1924.

Thurstone, L. L. "A Comparative Study of Clerical Tests," *Public Personnel Studies*, Vol. 1, Nos. 2–6, November, December, 1923.

Toops, H. A. "Tests for Vocational Guidance of Children Thirteen to Sixteen," *Teachers College, Columbia University, Contributions to Education*, No. 136. New York: Bureau of Publications, Teachers College, Columbia University, 1923. 159 p.

"Information and Data Regarding Tests Previously Published (Junior Clerk)," *Public Personnel Studies*, 5:240–41, November, 1927.

"Partially Standardized Tests for Junior Clerk," *Public Personnel Studies*, 3:346–72, December, 1925.

"Partially Standardized Tests for Senior Clerk," *Public Personnel Studies*, 5:144–58, July, 1927.

"Standardized Tests of Alphabetical Filing and Ability to Understand and Follow Written Directions," *Public Personnel Studies*, 5:80–88, April, 1927.

"Suggested Tests for Addressograph Operator," *Public Personnel Studies*, 6:197–201, September, 1928.

"Suggested Tests for Senior Account Clerk," *Public Personnel Studies*, 4:166–75, May, 1926.

VII. General

Blackstone, E. G. "Research Studies in Commercial Education," *University of Iowa Monographs in Education*, Series 1, No. 11. Iowa City: University of Iowa, 1929.

Blackstone, E. G., et al. "Research Studies in Commercial Education," *University of Iowa Monographs in Education*, Series 1, No. 7. Iowa City: University of Iowa, 1926. 160 p.

Cody, Sherwin. *Commercial Tests and How to Use Them*. Yonkers, New York: World Book Company, 1920. 216 p.

Filer, H. A. and O'Rourke, L. J. "Progress in Civil Service Tests," *Journal of Personnel Research*, 1:484–520, March, 1923.

Grizzell, E. D. "The Value of Prognostic Tests to Determine Potential Capacities with Special Application to Business Activities," *Eleventh Annual Schoolmen's Week Proceedings*. Philadelphia: University of Pennsylvania, 1924, p. 274–81.

Link, H. C. *Employment Psychology*. New York: The Macmillan Company, 1919. 440 p.

Ream, M. J. "A Social Relations Test," *Journal of Applied Psychology*, 6:69–73, March, 1922.

Thurstone, L. L. "A Comparative Study of Clerical Tests," *Public Personnel Studies*, Vol. 1, Nos. 2–6, November, December, 1923.

Whitman, A. D. "New-Type Testing in Commercial Subjects," *Fifteenth Annual Schoolmen's Week Proceedings*. Philadelphia: University of Pennsylvania, 1928, p. 560–64.

"A Method of Rating the History and Achievements of Applicants for Positions," *Public Personnel Studies*, 3:202–9, July, 1925.

CHAPTER XIII

HEALTH AND PHYSICAL EDUCATION

Introduction.—Comparatively little has been done by way of constructing standardized tests to measure health knowledge or habits, but a great deal has been done along many diverse lines of physical measurements. Long before standard tests were thought of in connection with academic school subjects, more or less standardized physical measurements were being made. The field covered by them is so broad that a whole volume could easily be devoted to it and scores, if not hundreds, of tests, measurements, indices, and so forth, discussed. In this chapter the writer will attempt only to describe very briefly a few of the different measuring instruments and methods in this field so as to indicate its scope rather than to list all of the best. Furthermore, it seems impracticable to attempt to divide the field into parts since there is such great overlapping among the methods of measuring. Certain divisions might be made such as health, anthropometric and physiological measurement, and the measurement of strength and athletic proficiency, but in view of the fact just stated, it does not seem particularly helpful to do so.

It is perhaps needless to say that few of the measures obtained are secured by the use of pencil and paper tests. They are rather measures of physical dimensions, of quality of performance including height or distance, time, and perhaps other factors, and so forth. Among the more common types may be mentioned measurement of pupils' knowledge of the various factors that contribute to health, of height, weight, strength of grip, lung capacity, heart action, ability to perform various stunts or athletic events, of condition of body organs, such as the eyes, ears, teeth, lungs, feet, and so forth, and of motor skill and coördination. In comparatively few secondary schools are extensive programs of this sort being carried on, and even in most of those that are doing much along this line little use is being made of the data collected.

There is, therefore, place for more activity both in actual measurement and in interpreting the results of measurement.

Health Knowledge Test
A. I. Gates and Ruth Strang (1925)
Complete Series; Form 1

Although little has been done by way of measuring health knowledge, this test deserves high rank. The complete series contains 520 multiple-answer exercises classified by topics, and Form 1 sixty-four selected from the longer list. It was apparently intended that other similar tests be made, but they have not yet appeared. Although mostly used in the elementary school, this is also designed for use throughout the high school. The basis of selecting the content was an analysis of twenty courses in health used in rural and city schools, and of fourteen of the most widely used texts. The 744 exercises constructed as a result of this study were tried out with elementary and high-school pupils, and college students, and criticized by sixteen specialists in psychology and the various subject-matter fields dealt with. Of those retained, ninety-eight deal with food, seventy-eight with disease, and the others in decreasing numbers with twenty-eight other topics. Forty minutes is the time for Form 1. Average scores for a typical New York school are given as follows:

Grade	*Low VII*	*High VII*	*Low VIII*	*High VIII*
Score	39	42	45	48

Bureau of Publications. Form 1, specimen set 15¢; $3.00 per 100. Complete series $1.00 per copy.

Reference: Gates, A. I. and Strang, Ruth. "A Test in Health Knowledge," *Teachers College Record*, 26:867–80, June, 1925.

Health Test for Junior-Senior High School
H. C. Pryor (1929)
Forms A, B

Part I has about seventy-five true-false statements and Part II twenty-nine multiple-answer exercises. The items cover the

same general field as those in the Gates-Strang test, including habits of eating, drinking, sleeping and personal cleanliness, choice of foods, dealing with diseases, and so forth. No time limit is stated, but how long each pupil takes is recorded. Twenty-five minutes seem to be ample. The coefficient of reliability is only .66. Norms based on about five hundred seventh-grade pupils and a much smaller number in each of the others are:

	Grade				
	VII	*VIII*	*IX*	*X*	*College*
Third quartile	59	59	67	69	79
Median	48	53	59	63	74
First quartile	40	44	51	52	67

H. C. Pryor. Sample set 10¢.

Scales for Health Habits, Attitudes, and Knowledge
T. D. Wood and M. O. Lerrigo (1927)
Scales IV, V, VI [1]

Scale IV deals with health habits, attitudes, and knowledge appropriate for pupils completing the ninth grade, V for those completing the twelfth, and VI with those for adults. Each is divided into three main divisions headed: The Healthy Organism, The Healthy Personality, and The Healthy Home and Community. Each of these is subdivided into several sections containing a number of statements of habits, attitudes, and knowledge grouped according to whether they are new or changed or repeated from scales for younger children, and also according to which one of the three classes—habits and skills, attitudes, or knowledge—they fall under. Scales IV and V each contain several hundred such statements, and Scale VI somewhat less. No provision is made for computing numerical scores, but it is apparently intended that the scales be used more as objectives or ideals to be held before pupils.

Public School Publishing Company. Book containing all scales and accompanying discussion, $2.00.

[1] There are also Scales I, II, and III, for kindergarten, end of the third and of the sixth grade, respectively.

Reference: Wood, T. D. and Lerrigo, M. O. *Health Behavior*. Bloomington, Illinois: Public School Publishing Company, 1927. 150 p.

Weight–Height–Age Tables for Children of School Age

T. D. Wood and B. T. Baldwin

Tables for Boys, for Girls

These two tables, which will not be reproduced here, are probably the best available standards for comparing the weight, height, and age of school children. Beginning with norms for primary children they extend far enough up to include most high-school pupils. They do not go beyond nineteen years for boys and eighteen for girls, nor beyond the height of seventy-four inches for boys and seventy-one for girls. In the actual use of such tables there is no generally agreed upon point below which pupils are considered seriously underweight or above which overweight. Sometimes 7 and sometimes 10 per cent below the norm is considered as indicating definite malnutrition, whereas another interpretation sometimes employed is that ten pounds below should be regarded as the danger point. In most instances comparatively little attention has been paid to overweight. This is unfortunate, for although overweight does not commonly indicate as dangerous a condition as does underweight, yet it should receive the proper amount of attention.

These tables are not for sale in the ordinary sense of the word, but may be found in numerous publications.

Reference: Baldwin, B. T. "The Use and Abuse of Weight-Height Tables as Indices of Health and Nutrition," *Journal of the American Medical Association*, 82:1–4, January 5, 1924.

Physical Test of a Man

D. A. Sargent (1921)

This, in common with the tables just described, is not offered for sale by any publisher, but is rather a simple physical index which has been described by Sargent and may be easily secured by following his directions. These provide that the individual to be tested should stand under a piece of stiff cardboard held from ten to twenty or more inches above his head. He then bends

forward, flexing trunk, knees, and ankles, and jumps upward, straightening and endeavoring to lift his head as high as possible. This is continued until the greatest height at which he can just touch the piece of cardboard is determined. The index is then determined by adding his weight in pounds to the height of the cardboard in inches, and dividing by his height in inches. Unfortunately no satisfactory norms for this have been established. In actually applying this test many variations have been introduced, some even involving time. Thus although the idea has received considerable approval, there is much disagreement as to the details.

References: Sargent, D. A. "The Physical Test of a Man," *School and Society*, 13:128–35, January 29, 1921. *Also in: American Physical Education Review*, 26:188–94, April, 1921.

Sargent, L. W. "Some Observations on the Sargent Test of Neuro-Muscular Efficiency," *American Physical Education Review*, 29:47–59, February, 1924.

Athletic Badge Tests

Playground and Recreational Association of America (1923)

Boys, Girls; Tests 1, 2, 3 for each

Each consists of four stunts or athletic events.[2] Considerable choice is allowed individuals taking the different tests; the events among which they may choose, however, are similar, so that in a complete test each individual must perform activities of different types. The events listed include dashes of various lengths, basketball and indoor-baseball throws, volley ball and tennis serves, balancing, potato and Indian club races, and so forth. It is intended that an individual shall begin with the first test and when he succeeds in passing the events in it, try those in the second, and after succeeding there, those in the third. Provision is made for a series of badges to be given pupils who complete the three tests.

It appears that these tests have been more commonly used in elementary than in high school, but they are intended for both, and at least some of them are difficult enough for high-school pupils who are not athletically inclined.

[2] An earlier edition of the same tests, issued in 1915, differed somewhat from the one described here.

Playground and Recreational Association of America. 5¢ per copy.

Reference: "Athletic Badge Tests for Boys and Girls," *U. S. Bureau of Education Physical Education Series* No. 2. Washington: Government Printing Office, 1923. 17 p.

Physical Achievement Tests for Girls and Women
Agnes R. Wayman et al. (1924)

These tests consist of four groups dealing with track and field events, stunts, games, and miscellaneous. The first group includes the fifty-yard dash, basketball throw, high jump and target throw, the second various stunts classified under the heads of balancing, agility, and strength, the third practice and actual playing of basketball, baseball, soccer and hockey, and the last walking, swimming, and tennis. For each event there is a graduated scale of points according to proficiency.

Woman's Division, National Amateur Athletic Federation of America.

Physical Capacity Tests
F. R. Rogers (1925)
Tests 1, 2, 3, 4, 5, 6, 7

This is probably one of the best series of tests dealing with vital capacity and the strength of the larger muscle groups. No. 1 deals with forced lung capacity, 2 and 3 with strength of grip of the two hands, respectively, 4 with strength of back, 5 with strength of legs, 6 and 7 with strength of arms, the former being "push-ups" and the latter "pull-ups." The whole series of tests should be given an individual in about ten minutes. Provision is made for combining scores on the seven tests into a composite strength index. It has been shown that this index has a high correlation with measures of athletic proficiency both outdoor and indoor. Its reliability is also high, the coefficient being .94.

In addition to the strength index Rogers also suggests a physical fitness index in which he corrects the former one for age and weight, and thus determines how each individual compares with the norm for others with whom he is comparable.

These tests also do not need to be purchased, but can be given by anyone possessing the necessary equipment and familiarizing himself with the directions.

Reference: Rogers, F. R. "Physical Capacity Tests in the Administration of Physical Education," *Teachers College, Columbia University, Contributions to Education*, No. 173. New York: Bureau of Publications, Teachers College, Columbia University, 1925. 93 p.

A Scale of Motor Ability Tests

D. K. Brace (1927)

Tests 1–20

This series of tests calls for the performance of such activities as walking in a straight line, placing the heel of one foot against the toe of the other, jumping into the air and clapping the feet together, lying flat on the floor with arms folded across the chest and raising the trunk to a sitting position without raising the feet, jumping and making a full turn to the left, standing kick so that toes at least come up level with the shoulders, sitting down cross-legged and arising again with arms folded on chest and without moving feet, and so forth. Each attempt is scored as a success or failure, and the score is the number of tests passed. Provision is made for changing this score into a so-called scale score for ages that include most high-school pupils.

A. S. Barnes and Company. 50¢ per copy, $10.00 per 25, $15.00 per 50, $25.00 per 100; scoring blank 10¢ per copy.

Reference: Brace, D. K. *Measuring Motor Ability: A Scale of Motor Ability Tests*. New York: A. S. Barnes and Company, 1927. 138 p.

In addition to the measurements of the sort suggested by these tests, there are many others not represented by particular tests which are frequently used. In connection with most of these it is important that they be carried out according to the proper directions, and thus in a sense they may be considered standardized tests. One of these is the measurement of chest girth and capacity, for which Seaver's[8] directions are probably as good as any. The

[8] Seaver, J. W. *Anthropometry and Physical Examination*. New Haven, 1890, p. 50–53.

standard norms for chest capacity are those of Smedley,[4] given by Whipple [5] and others. Another physiological measurement frequently made is that of pulse rate. One method of interpreting the results in terms of general physical efficiency has been suggested by Foster.[6] The chief objection to this is that his interpretation is not definite enough. Another and probably much better one is that of Schneider,[7] who has suggested an index to measure the cardiovascular system, based on heart rate and blood pressure taken under different conditions. The chief variation is that these measurements are made after a period of rest and after exercise, and also while the subject assumes different postures.

BIBLIOGRAPHY

Baldwin, B. T. "A Measuring Scale for Physical Growth and Physiological Age," *Fifteenth Yearbook of the National Society for the Study of Education*, Part I. Bloomington, Illinois: Public School Publishing Company, 1916, Chapter I.

Bovard, J. F. and Cozens, F. W. "Tests and Measurements in Physical Education 1861–1925," *University of Oregon Publication, Physical Education Series*, Vol. 1, No. 1. Eugene: University of Oregon, 1926. 94 p.

Brownell, C. L. "A Scale for Measuring the Antero-Posterior Posture of Ninth Grade Boys," *Teachers College, Columbia University, Contributions to Education*, No. 325. New York: Bureau of Publications, Teachers College, Columbia University, 1928. 62 p.

Franzen, Raymond. "Physical Measures of Growth and Nutrition," *School Health Research Monographs*, No. 2. New York: American Child Health Association, 1929. 138 p.

Franzen, Raymond et al. "Health Education Tests," *School Health Research Monographs*, No. 1. New York: American Child Health Association, 1929. 70 p.

Hanmer, L. F. "Physical Training and Play," *The Gary Public Schools*. New York: General Education Board, 1919. 35 p.

Payne, E. G., et al. *Education in Health*. New York: Lyons and Carnahan, 1921, p. 229–43.

[4] Smedley, F. "Report of Department Child Study and Pedagogic Investigation," *Forty-sixth Annual Report, Board of Education, Chicago*, 1899–1900, 1900–1901.

[5] Whipple, G. M. *Manual of Mental and Physical Tests*, Part I. Baltimore: Warwick and York, 1914, p. 70–74.

[6] Foster, W. C. "Test of Physical Efficiency," *American Physical Education Review*, 19:632–36, November, 1914.

[7] Schneider, E. C. "A Cardiovascular Rating as a Measure of Physical Fatigue and Efficiency," *Journal of American Medical Association*, 74:1507–10, May 29, 1920.

Payne, E. G. "The Measurement of Social Values—A Scale for Measuring Health Practices," *Contributions to Education*, Vol. 1. Yonkers, New York: World Book Company, 1924, Chapter XV.

————. "A Scale for Measuring Habits and Practices in Health and Accident Prevention," *School and Society*, 17:25–27, January 6, 1923.

Rapeer, L. W. "Minimum Essentials of Physical Education and a Scale for Measuring Results of Physical Education," *Sixteenth Yearbook of the National Society for the Study of Education*, Part I. Bloomington, Illinois: Public School Publishing Company, 1917, Chapter XI.

Schwegler, R. A. and Engelhardt, J. L. "A Test of Physical Efficiency: The Correlation between Results Therefrom and Results from Tests of Mental Efficiency," *American Physical Education Review*, 29:501–5, November, 1924.

Smith, H. L. and Wright, W. W. "Health and Physical Education," *Tests and Measurements*. New York: Silver, Burdett and Company, 1928, Chapter XVII.

Sorden, H. L. "What is Your Health-Beauty Score?" *Child Welfare Magazine*, 19:422–24, April, 1925.

Symonds, P. M. "Tests for Physical Education," *Measurement in Secondary Education*. New York: The Macmillan Company, 1927, Chapter X.

Way, A. P. and Atkinson, R. K. "Tests in Physical Education," *Contributions to Education*, Vol. 1. Yonkers, New York: World Book Company, 1924, Chapter XXIX.

Wayman, A. R. "A Scheme for Testing and Scoring the Physical Efficiency of College Girls," *American Physical Education Review*, 28:415–20, November, 1923.

Wilson, G. M. and Hoke, K. J. "Measurement in Physical Education," *How to Measure*, Revised and Enlarged. New York: The Macmillan Company, 1928, Chapter XIV.

CHAPTER XIV

MISCELLANEOUS SUBJECTS

Introduction.—In this chapter will be grouped a few subjects in which not enough has been done in the way of making standardized tests to justify separate chapters, and that do not belong in any of the subjects groups discussed in preceding chapters. Some of them are more or less regular high-school subjects, whereas others are not, but possess interest and value for high-school teachers.

I. Agriculture

Agriculture is undoubtedly the most commonly taught elementary or high-school subject in which there are no real standardized tests available. Probably the chief reason for this is that most courses in this subject are adapted to the needs of the community in which they are taught and that these needs vary so much from place to place that it is very difficult to construct any tests that fit the content of the courses taught throughout a large extent of territory. If tests are limited to the few common items, they are so short and cover so little of the subject matter as to be of little help. If they are limited to particular topics such, for example, as corn production, dairying, poultry raising, and so forth, the difficulty is encountered that the best methods of carrying on such activities depend largely on local conditions of various sorts and therefore are not uniform.

Even though there is available no thoroughly standardized series of tests in agriculture, one series, the only one so far as the writer knows that approaches standardization, will be described below.

National Agricultural Tests

C. E. Myers, C. E. Cronemeyer, A. D. Wilson, and W. M. Hendricks (1924)

Vegetable Gardening, Poultry, Farm Crops; Forms A, B of each

This series has been in process of development for at least

five or six years and is not yet completed, there being three tests still in mimeographed form in addition to the three mentioned above. The vegetable gardening test consists of one hundred true-false statements embodying miscellaneous information in that field, eighteen matching items dealing with varieties of vegetables, and an equal number with the identification of pictures of vegetables. The poultry information test calls for identifying fifteen parts of a pictured fowl, and responding to 135 true-false statements, multiple-answer exercises, and pairs of matching items composed of terms and their definitions. That on farm crops includes 125 true-false and multiple-answer statements, and thirty definitions to be matched with the proper terms. The time limit for each test is thirty minutes.

The items included in these tests were selected on the basis of a study of a number of widely-used texts, bulletins, and other similar material, and of lists of questions prepared by senior students in "Methods of Teaching Agriculture." The original list was criticized by about thirty experienced teachers and agricultural specialists, then tried out before the two final forms of each test were constructed. The median scores for various groups based upon from five hundred to one thousand cases are as follows:

	Poultry	
	Form A	*Form B*
Experienced teachers of agriculture	125	126
College seniors who have studied poultry one semester	114	116
College seniors who have studied poultry three weeks	88	89
Students who have carried poultry projects	93	100
High-school seniors who have studied poultry	75	81
High-school sophomores and juniors who have studied poultry	60	64
Entering high-school freshmen	30	30

For high-school pupils,

$$r = .87,\ \text{P.E.}_{\text{meas.}} = 6,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .09,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .24.$$

Rural Education Department, State College, Pennsylvania. Sample set 15¢; 50¢ per 8, $1.00 per 24, $2.00 per 50.

II. Mechanics and Engineering

In dealing with tests of mechanical and engineering aptitude and ability it is difficult to know where to draw the line between them and manual training on the one hand, and between them and vocational guidance instruments on the other. The effort has been made, however, to include in this section tests that are primarily prognostic in their function and that do not deal with the content of courses commonly offered in high school. Furthermore, they are differentiated from many vocational-guidance instruments in being rather definitely tests as contrasted with scales, questionnaires, information blanks, and so forth. Most of those included here look forward to the courses to be carried or the vocations entered by pupils after leaving the secondary school, although some may be used to aid in the selection of subjects therein, especially in connection with technical work in large city schools.

There has been considerable activity in this field apart from that having any connection with the schools as such. Probably the outstanding example of this is the work done in the army.[1] Among the tests of this sort employed in the army may be named those for automatic screw machine operators, auto repairers, die sinkers, electricians, machinists, typewriter repairers, and so on. Some of the army tests are oral, consisting of questions to be asked in an interview, some are written, of the ordinary pencil and paper type, and others are actual performance tests requiring the testee to construct some object or perform a specified bit of work under standard conditions.

Assembling Tests of General Mechanical Ability

J. L. Stenquist (1918)

Tests I, II

This test consists of a box, approximately six inches by two feet, divided into ten compartments, each of which contains the parts of a mechanical contrivance. The devices included were selected as being of a mechanical nature, but yet coming within the common experience of most persons. They include a bicycle

[1] Chapman, J. C. and Chapman, Daisy. *Trade Tests.* New York: Henry Holt and Company, 1921. 435 p.

bell, an electric pushbutton, a mouse trap, an expansion nut, a simple lock, and so forth. Pupils are given thirty minutes within which to assemble the ten articles. These two series were arranged after six years of experimental work with other series that contained a number of the same contrivances. The devices were carefully scaled and apparently are arranged in order of increasing difficulty. Test scores have a coefficient of reliability of about .70, a correlation with intelligence test scores of only about .30 and with school marks in more or less mechanical subjects of about .80. These tests are probably the most satisfactory measures of general mechanical ability available, but because of their high cost are not widely used.

Two methods of scoring are suggested. In one ten points are allowed for each device, with partial credit according to how much of it has been correctly assembled. This may be said to be the approved scoring method and the norms below are given in terms thereof. The other or short method is merely to count the number correctly assembled. If this is done all devices assembled nearly enough correctly to deserve scores of eight or nine as well as ten are counted correct. A time bonus of one-half point for each minute less than thirty is also provided. Norms according to age are as follows:

Age	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
Fifteen	1.0	1.7	3.2	4.6	6.2	7.2	7.9
Fourteen	1.0	1.4	2.4	4.3	6.0	7.4	8.0
Thirteen	1.0	1.5	2.5	3.9	5.3	6.6	7.7
Twelve	.7	1.0	1.8	2.9	3.8	5.2	6.8

These norms are the averages of results from the two series. It will be seen that there are some irregularities in that the figures for pupils of a given age are sometimes above those for older pupils, and that on the whole there is a comparatively small increase from year to year.

C. H. Stoelting Company. $12.50 per set; scoring blanks 80¢ per 100.

References: Stenquist, J. L. "Measurements of Mechanical Ability," *Teachers College, Columbia University, Contributions to Education*, No. 130.

New York: Bureau of Publications, Teachers College, Columbia University, 1923. 101 p.

Toops, H. A. "Tests for Vocational Guidance of Children Thirteen to Sixteen," *Teachers College, Columbia University, Contributions to Education*, No. 136. New York: Bureau of Publications, Teachers College, Columbia University, 1923. 159 p.

Yerkes, R. M. (Edited by). "Psychological Examining in the United States Army," *Memoirs of the National Academy of Sciences*, Vol. 15. Washington, D. C.: Government Printing Office, 1921, p. 91, 147–49, 267, 321–23.

Mechanical Aptitude Tests

J. L. Stenquist (1921)

Tests I, II

Because of the high cost and comparative difficulty of employing the same author's Assembling Tests, he prepared these pencil and paper tests also. They are not duplicate forms, but are to be used together. The first consists of six exercises in each of which are represented three or four pairs of sets of five mechanical objects each. One in each set is to be paired with an associated object in the other set. For example, in one set are five types of hammers and mallets, and in the corresponding set five different objects commonly driven. Test II contains only three exercises. The first is similar to those in Test I except that there are nineteen objects in each set. The second contains representations of four mechanical objects with from five to eleven questions about each calling for identification of parts, knowledge of movements and results therefrom, and so forth. In Exercise 3 are two figures and also representations of a number of parts. There are questions to be answered about the figures and also the parts are to be matched with the proper portions of the figures. The two tests together call for 173 responses. They deal with devices as mechanical in nature as possible, but at the same time of rather general interest. The working time on the first test is forty-five minutes and on the second fifty.

$$r = .80,\ \text{P.E.}_{\text{meas.}} = 3,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .10,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .30.$$

Their correlation with the Assembling Tests has been reported as below .70 and also as above .80, with estimates of general mechanical aptitude as .84, and with teachers' marks in appropriate

courses as .60. With intelligence tests their correlation is low. Stenquist has shown that the aptitude they measure is not primarily a function of definite training. On the whole the measures yielded are less valid for the indicated purpose than those from the Assembling Tests. The norms for these tests are given in terms of ages as follows:

Age	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
Test I							
Fifteen	21	26	34	42	54	69	78
Fourteen	20	25	32	40	52	64	73
Thirteen	19	24	31	38	47	56	63
Twelve	18	22	29	35	42	50	56
Test II							
Fifteen	17	21	28	36	44	52	55
Fourteen	16	20	27	35	43	50	54
Thirteen	15	19	26	34	42	49	52
Twelve	12	16	24	31	39	47	50

World Book Company. Specimen set 30¢; $1.50 per 25; manual 15¢.

References: Stenquist, J. L. "Measurements of Mechanical Ability," *Teachers College, Columbia University, Contributions to Education*, No. 130. New York: Bureau of Publications, Teachers College, Columbia University, 1923. 101 p.

Bell, J. C. "Mechanical Aptitude and Intelligence," *Contributions to Education*, Vol. 1. Yonkers, New York: World Book Company, 1924, p. 270–82.

Test for Mechanical Ability

T. W. MacQuarrie (1925)

This test is composed of seven subtests dealing, respectively, with tracing, tapping, dotting, copying, location, blocks, and pursuit. In other words, it attempts to measure certain motor and other abilities rather than mechanical knowledge. Each subtest is preceded by a similar but shorter one to be used as a practice exercise, thus insuring familiarity with what is to be done. The subtitle of the test "A Simple Group Performance Test for the Use of School Counselors and Personnel Managers" indicates its designed function. The writer states that evidence from other similar tests shows that about thirty minutes of testing will yield as

accurate an estimate of mechanical ability as can be made by a teacher after a year's close observation. The actual working time is less than fifteen minutes in periods of ten seconds or more, but a total of about thirty is required.

$$r = .90,\ \text{P.E.}_{\text{meas.}} = 3,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .07,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .21.$$

A reported tryout with a comparatively few pupils indicates low correlation between test scores and intelligence quotients, .66 with the rating of a rather complicated mechanical project in electrical construction, .72 with the time required in carrying out the project, .79 with the average of the rating and the time, and from .48 to .80 with teachers' estimates of mechanical ability. The following age norms are reported:

Age	-1σ *	*Mean*	$+1\sigma$ *
20	52	68	84
19	51	67	83
18	49	65	81
17	47	63	79
16	45	60	75
15	43	57	71
14	40	53	66
13	37	49	61
12	33	44	55

* -1σ denotes one standard deviation below the mean, or approximately the 16th percentile, and $+1\sigma$ one standard deviation above the mean, or about the 84th percentile.

Research Service Company. $6.00 per 100.

References: Leonard, R. S. and Horning, S. D. "A Study of the Adequacy of Tests for Mechanical Ability as a Basis for Judging Mechanics Arts Students," *Educational Research Bulletin of the Pasadena City Schools*, Vol. 4, Nos. 7, 8. Pasadena, California: Bureau of Research and Service, 1926, p. 4–8.

Stein, M. L. "A Trial with Criteria of the MacQuarrie Test of Mechanical Ability," *Journal of Applied Psychology*, 11:391–93, October, 1927.

MacQuarrie, T. W. "A Measure of Mechanical Ability." Los Angeles: Associated Students Store, University of Southern California, 1925.

———. "A Mechanical Ability Test," *Journal of Personnel Research*, 5:329–37, January, 1927.

Mechanical Aptitude Test—Junior Grade

L. J. O'Rourke (1926)

This test, which is one of O'Rourke's Series of Aids in Placement and Guidance, consists of two parts. In the first are twelve sets of from three to twelve representations of tools, machines, parts, and other mechanical objects, grouped in six pairs. There are two exercises on each pair; the first calls for matching one object from each set with the proper part from the other set, as, for example, a screw driver with a screw, a mallet with a chisel, a tire pump with an air pressure gauge, a connecting rod with a crank shaft, a T-square with a triangle, and so on. The second exercise following each pair of groups asks for responses indicating which of the tools shown should be used in connection with each of from three to nine operations or purposes, such as to keep a gate closed, to thread a hole in metal, to prevent oil from being sucked into a cylinder, and so on. Part 2 consists of sixty multiple-answer items dealing with the same type of content. Part 1 is preceded by a fore-exercise containing two groups of objects and two exercises similar to those in the test itself. Six minutes are allowed to study the fore-exercise, thirty for Part 1, and twenty-five for Part 2, with a suggestion that Part 2 may be given at a second sitting if desired. The coefficient of reliability is given as .94.

Educational and Personnel Publishing Company. 10¢ per copy, $5.00 per 100.

Electrical Inclination Test for Vocational Guidance

N. W. Ruth (1928)

This test is intended to be both diagnostic and prognostic, measuring both interest in electricity and acquired information concerning it. Following a sample exercise are six tests. The first four of these consist of three sets of five objects each, and three sets of five related objects, between which the proper connections are to be made. For example, the proper kind of electric bulb is to be matched with an ordinary socket, another kind with an automobile headlight, another with a flashlight, and so on. Each related pair of groups deals with objects closely associated, that

is, all used for very similar purposes. Test 5 consists of eighteen true-false statements and Test 6 of fifteen multiple-answer statements dealing with various facts concerning electricity, its use and operation, and in a few instances its history.

C. H. Stoelting Company. $4.50 per 25, $13.50 per 100; manual 20¢; scoring key 15¢.

Detroit Mechanical Aptitudes Examination

H. J. Baker and A. C. Crockett (1928)

Boys, Girls

Each of the two tests contains eight subtests. Subtests 1, 4, 7, and 8 are similar in form, but differ in content for boys and girls, whereas the other four are exactly the same. The first requires the matching of forty tools, such as an ax, a draw knife, a hack saw, a putty knife, and so forth for boys, and a butcher knife, a hammer, a pickle fork, and a trowel for girls, with their proper names. Subtest 2 requires that a line be drawn between a pair of very much curved and broken lines about one-eighth of an inch apart without touching either. The third presents sixteen sets of from three to eight lines, squares, angles, and other figures and requires that those in each set be labelled in order of size. The fourth, in multiple-answer form, deals with the meaning or use of such articles as lock washer, acetylene torch, and mitre box for boys, and sausage, solder, and carburetor for girls. Subtest 5 consists of over one hundred representations of screws, nuts, bolts, and washers, and requires that the number under each be put in the proper place to indicate which it is. The next one consists of twenty geometrical figures each followed by five others of which some two could be combined to form the first. These, of course, are to be indicated. The seventh contains reproductions of such objects as a plane, a vise, and so forth for boys, a food chopper and a pressure cooker for girls, and requires that forty numbered parts be matched with their proper names or statements of functions. The last subtest consist of six problems, those for boys dealing with pulleys, and those for girls with stitches. In the former case the direction and relative speed of the various parts of the figures must be indicated, and in the latter the order in which made and the direction of the needle. The total working time is twenty-

eight minutes, but the directions are so long that probably at least forty-five will be required to give the test.

The selection of test items was intended to reproduce as closely as possible actual vocational situations. Knowledge about tools, motor skill, and visual acuity are supposed to be measured. Intercorrelations among the different subtests range from about .20 to .80 and of each with the composite of the other seven from about .40 to .70. In general those for the boys' tests are several points higher than for the girls'. The correlation of boys' scores with shop teachers' ratings for mechanical aptitude is about .64, and that with intelligence for both tests probably about .40. The coefficient of reliability is reported as .76 for the boys' test and .87 for the girls'. Age norms are given as follows, apparently based on several thousand cases.

Year	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
Boys	100	110	119	127	134	143	157	174	195
Girls	79	87	96	103	109	117	129	144	161

Public School Publishing Company. Sample set each 15¢; 4¢ per copy, $3.00 per 100.

Reference: Baker, H. J. "A Mechanical Aptitude Test," *Detroit Education Bulletin*, 12:5–6, January, 1929.

Vocational Guidance Tests

L. L. Thurstone (1919)

Algebra, Arithmetic, Geometry, Physics, Technical Information

This series is sometimes also known as the Engineering Aptitudes Tests. Each of the first four tests includes from nine up to twenty-five problems and in some cases other exercises, and the Technical Information Test one hundred multiple-answer items. Thirty minutes are allowed for each. Although these tests were originally given to about seven thousand engineering students in forty-three colleges, it does not appear that they were successful enough for their intended purpose to result in their use for this purpose becoming very common. Correlations between the scores on the separate tests and first-year scholarship average between .30 and .40, which is not high enough to be of very much value in prediction. Figures are not given for the combined scores from

the whole series, but no doubt such scores would yield considerably higher correlations. The intercorrelations of the tests range from .14 up to .56, averaging less than .40. Although the tests do not appear to have great prognostic value, they have continued to receive some use as ordinary tests of the subjects dealt with. This is especially true of those in algebra, geometry, and physics. However, it does not appear that they should be ranked among the best tests now available.

World Book Company. Specimen set 40¢; $1.00 per 25; manual 20¢.

III. General Survey Tests

Although most of the so-called general survey or general achievement tests covering a number of subjects are designed for use in the elementary school, there are three or four that may well be employed in high school. These are, it is true, largely if not entirely intended for use in connection with college entrance, but since they cover the more important subjects commonly taught in high school, it seems that they may likewise appropriately be employed in the senior year of high school at least, and perhaps below.

There is one marked difference between such tests for the secondary school and similar tests for the elementary school. The latter commonly include a number of so-called fundamental elementary subjects on the assumption, which is usually true, that all elementary pupils have studied them. No such assumption covering any considerable number of high-school subjects is justified; therefore some of the tests to be described in this section have either provided certain sections that all pupils are to take and others to be taken only by those pupils who have studied the subjects dealt with, or are printed in separate parts so that each pupil takes only those that fit his case.

Iowa High School Content Examination
G. M. Ruch et al. (1923)
Forms A, B, A–1, B–1

This examination is Part III of the Iowa Entrance Examination, the other parts being the Thorndike Intelligence Examina-

tion [2] and the Iowa Comprehension Tests.[3] Its four sections deal with English literature, mathematics, science, and history and social science. In Forms A and B there are 110 elements on English literature, 75 on mathematics, 100 on science, and 115 on history and social science, or 400 in all. Forms A–1 and B–1 are shorter, the corresponding numbers being 75, 50, 50, and 75, a total of 250. All the exercises are multiple-answer, with five possible answers. For Forms A and B the time allowed is twenty minutes per section, or eighty in all, for Forms A–1 and B–1, ten minutes for the third section and fifteen for each of the others, making a total of fifty-five minutes.

The items included appear to be such that the use of this examination should not be limited to Iowa. Indeed, it has been used in several other states with good results. A correlation of about .50 between scores on this test and first-semester marks has been found at the University of Iowa. Average data on the reliability of the separate sections are given in the first line of the table below:

Form	r	$P.E._{meas.}$	$\frac{P.E._{meas.}}{M}$	$\frac{P.E._{meas.}}{\sigma}$
A and B (sections)	.90	3	.08	.22
A and B	.95	7	.04	.15
A–1 and B–1	.95	5	.04	.14

Mean scores for the separate sections, based on April testing of about two thousand high-school seniors, are:

	Form		
	B	*A–1*	*B–1*
English	50	38	40
Mathematics	31	19	22
Science	36	13	20
Social science	55	36	37

In addition the following percentile norms for total scores have been published. These are based on the same high-school seniors

[2] *See* page 410.
[3] *See* page 138.

and more than a thousand university freshmen tested in September.

	Percentile						
	5	*10*	*25*	*50*	*75*	*90*	*95*
FORM B							
University	100	121	147	181	219	255	276
High school	85	102	132	167	206	242	264
FORM A–1							
University	57	68	87	110	134	158	173
High school	52	63	80	102	129	159	174
FORM B–1							
University	70	80	98	122	144	164	178
High school	63	75	94	116	141	162	177

Bureau of Educational Research and Service. 10¢ per copy; $8.00 per 100.

Reference: Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction.* Yonkers, New York: World Book Company, 1927, p. 202–4.

IOWA PLACEMENT EXAMINATIONS

C. E. Seashore and G. M. Ruch[4] (1924)

Chemistry—Aptitude and Training—G. D. Stoddard and J. Cornog

English—Aptitude and Training—M. F. Carpenter and G. D. Stoddard

Foreign Language—Aptitude—G. D. Stoddard and G. E. VanderBeke

French Training—G. E. VanderBeke and G. D. Stoddard

Spanish Training—G. E. VanderBeke and G. D. Stoddard

Mathematics—Aptitude and Training—G. D. Stoddard and E. W. Chittenden

Physics—Aptitude and Training—G. D. Stoddard and C. J. Lapp

Forms A, B, of each

Each test consists of four parts and requires from forty to

[4] Seashore and Ruch planned and directed the construction of the entire series, whereas the men named after the several subjects actually made the tests therein.

forty-five minutes of actual working time. The Chemistry Aptitude Test consists of twenty problems dealing with arithmetical relations that appear in chemistry, three paragraphs from chemistry textbooks followed by ten true-false statements based thereon, a longer selection upon which there are fifteen direct questions, and sixty true-false statements dealing with general chemical knowledge that may be acquired more or less incidentally.

The Chemistry Training Test contains forty-five true-false items dealing with fundamentals of chemical processes and fifty with manufacturing processes in the application of chemistry, and twelve fundamental chemical problems that involve rather simple arithmetic. Also it calls for the valences of twenty elements or ions, the formulae for ten compounds, the names of ten compounds for which formulae are given, and the completing and balancing of five equations.

The English Aptitude Test includes twenty sentences to be marked according to whether they comply with given grammatical rules or not, a selection taken from a college textbook, following which ten responses to multiple-answer exercises are called for, a selection of poetry followed by fifteen direct questions and a theme followed by twenty multiple-answer exercises that deal with its analysis.

The English Training Test calls for selecting twenty-five misspelled words out of a list of seventy-five, marking sixty examples according to whether they are punctuated correctly or not, sixty other examples with regard to grammatical errors, and forty-five others as to sentence structure.

The Foreign Language Aptitude Test begins with fifty English words and sentences of which the number, tense, degree, and so forth must be changed and derivatives formed from verbs. After this are forty Esperanto sentences each with an italicized word to be recognized and translated. Next thirty responses that measure ability to comprehend words and apply rules of grammar to Esperanto are called for, and finally forty more dealing with the translation of Esperanto into English, English into Esperanto, and the recognition of Esperanto parts of speech.

The French Training Test first asks for the English of sixty French words, then gives forty French sentences each of which

contains a mistake of grammar or idiom that is to be indicated, forty others in each of which the proper verb form must be selected from among four suggested, and finally three paragraphs in French followed by a total of twenty English questions to be answered in English.

The Spanish Training Test consists of fifty words for each of which the proper one of five meanings is to be chosen, forty sentences each containing a grammatical mistake, forty others in each of which a given verb is to be put in the proper form, and finally three paragraphs with questions similar to those in the French test.

The Mathematical Aptitude Test contains fifteen arithmetical and algebraic number series to each of which two terms are to be added, fifteen written exercises that deal with constructive imagination, twenty statements in each of which one or more mathematical facts are given and a conclusion drawn that is to be marked true or false, and a reading selection from calculus followed by fifteen direct questions.

In the Mathematical Training Test are twenty exercises in the fundamentals of arithmetic, twenty on operations and formal algebra, forty true-false statements on basic geometrical information, and fifteen algebraic reasoning problems involving rather simple arithmetic.

The Physics Aptitude Test has twenty-five problems dealing with arithmetic relations in elementary physics, three paragraphs from a college physics textbook each followed by ten true-false statements, five number series to each of which two additional terms are to be added, fifteen elements dealing with symbolic logic, mostly mathematical, in which the conclusions are to be marked true or false, and fifty true-false statements on common items of physics knowledge.

The Physics Training Test contains forty completion statements on fundamental information, fifty true-false ones on information and principles, twenty dealing with the completion of equations and statements and the names of fundamental laws or principles, and fifteen written problems in physics.

The items included in a tentative edition were chosen chiefly according to the judgment of supposedly competent persons who had had considerable teaching experience. Their choice was made

in the light of the following general principles arrived at by Seashore:

> (1) It [the test] will be devoted to a single subject or field of knowledge such as English, mathematics or chemistry; (2) it will differentiate between training in a subject and natural aptitude or fitness for that field of work; (3) it will be a departmental affair and will be given separately by each department in its immediate interests and needs; (4) it will serve as an introduction to the subject, being prepared partly with the purpose of reminding the student of the essential prerequisites for the course and indicating the general character of the activity that will be pursued in the course, and being so written from the point of view of the art of teaching that it shall constitute the most profitable exercise for the first two hours of the course; (5) this examination should give, at the end of two hours, as adequate information about the student's place and needs in the course as the instructor ordinarily acquires by the end of the first semester under the traditional methods of instruction; (6) the record of a general intelligence test may be used to supplement this examination, but that is not essential, as a series of placement tests will be more significant than a general intelligence test; (7) it will be prepared by or in responsible collaboration with a successful teacher and writer in the specific subject; (8) it will be given for a specific purpose and the results will be applied immediately in the organization of sections of the class on the basis of this objective information about the character of the preparation and the natural aptitude for the subject.

After the tentative edition had been tried out, items that appeared too hard, too easy, or for some other reason were unsatisfactory, were eliminated, a number of new ones added, and some mechanical improvements in scoring and form made. Considerable evidence has been accumulated having to do with the validity of these tests. Correlations with semester marks have been found to range as follows:

Chemistry aptitude	.23–.63	French training	.45–.65
Chemistry training	.25–.67	Spanish training	.48–.57
English aptitude	.25–.69	Mathematics aptitude	.16–.65
English training	.28–.67	Mathematics training	.34–.65
Foreign language aptitude	.36–.73	Physics aptitude	.28–.62
		Physics training	.57–.69

As would be expected, the use of both Forms A and B yields somewhat higher correlations, but combining scores on the training and aptitude tests gives less increased accuracy of prediction than would be expected. Composite scores on all the tests correlate about .75 with average academic success. The various training and

aptitude tests in the same subjects correlate with one another from about .50 to .70, and with those in different subjects from about .25 to .60. The aptitude tests average about .60 and the training tests about .52 with an intelligence test. Reliability data are as follows:

Test	r	$P.E._{mean}$	$\frac{P.E._{mean}}{M}$	$\frac{P.E._{mean}}{\sigma}$
Chemistry aptitude	.88	5	.09	.23
Chemistry training	.93	6	.09	.18
English aptitude	.82	3	.07	.29
English training	.90	8	.08	.21
Foreign language aptitude	.97	3	.04	.12
French training	.93	5	.08	.18
Spanish training	.82	5	.11	.29
Mathematics aptitude	.86	3	.11	.25
Mathematics training	.88	3	.09	.23
Physics aptitude	.83	8	.07	.28
Physics training	.85	8	.11	.26

The norms are based upon from one thousand to four thousand individuals from a number of different colleges in each case except in the case of the French and Spanish training tests. They are as follows:

Test	*First quartile*	*Median*	*Third quartile*
Chemistry aptitude	44	60	74
Chemistry training	47	71	97
English aptitude	33	40	46
English training	65	91	120
Foreign language aptitude	54	75	95
French training	49	67	88
Spanish training	33	41	53
Mathematics aptitude	21	29	37
Mathematics training	26	36	45
Physics aptitude	115	132	145
Physics training	46	69	92

A report of the use of the tests in engineering institutions based on about twice as many cases as the norms just given shows only two or three marked differences.

Test	Percentile						
	5	10	25	50	75	90	95
Chemistry aptitude	20	27	38	54	69	81	89
Chemistry training	16	25	45	69	95	117	130
English aptitude	24	28	34	41	47	52	56
English training	38	47	66	95	124	148	159
Foreign language aptitude	34	43	59	77	98	115	128
French training	23	30	42	63	88	115	128
Spanish training	19	24	33	41	54	73	77
Mathematics aptitude	9	12	17	26	35	42	47
Mathematics training	13	17	25	34	44	52	57
Physics aptitude	66	77	96	120	137	150	156
Physics training	20	29	47	69	93	111	123

Bureau of Educational Research and Service. Sample set 45¢; $3.50 per 100.

References: Hammond, H. P. and Stoddard, G. D. "A Study of Placement Examinations," *University of Iowa Studies in Education*, Vol. 4, No. 7. Iowa City: University of Iowa, 1928. 59 p.

Langlie, T. A. "The Iowa Placement Examinations at the University of Minnesota," *Journal of Engineering Education*, 17: 842–60, May, 1927.

———. "Analysis of the Iowa Placement Tests," *Journal of Applied Psychology*, 10:303–14, September, 1926.

Seashore, C. E. "College Placement Examinations," *School and Society*, 20:575–77, November 8, 1924.

Stoddard G. D. "Iowa Placement Examinations," *School and Society*, 24:212–16, August 14, 1926.

———. "Iowa Placement Examinations," *University of Iowa Studies in Education*, Vol. 3, Iowa City: University of Iowa, 1925. 103 p.

———. "Iowa Placement Examinations—A New Departure in Mental Measurement," *Psychological Monograph, University of Iowa Studies in Psychology*, Vol. 12. Iowa City: University of Iowa.

North Carolina High School Senior Examination

M. R. Trabue et al. (1927)

Editions of 1927, 1928, 1929, 1930

The first or 1927 edition of this examination consists of Sections A to H as follows:

A—English literature and forms
B—Comprehension of reading
C—Mental agility
D—History, American and general
E—Modern times in civics
F—General science
G—Mathematics
HL—Latin
HF—French

Most of these sections consist of twenty-five exercises, usually multiple-answer, each, but A and C have fifty apiece. The 1928 and later editions differ somewhat in form and content. Thus that of 1929 is divided into sections as follows:

A—Reading and study habits
B-1—Reading (literature)
B-2—Reading (history)
C—English
D—Mathematics
E—General science
F—American history
GL—Latin
GF—French
GA—Agriculture

Each of its sections contains twenty-five exercises. Each pupil is supposed to take Sections A to G of the 1927 edition or A to F of the 1929 and one of the others. The time for the complete examination is about an hour and a half.

Although this examination is particularly suited to work offered in North Carolina high schools, the later editions contain much less local matter than either of the others and is reasonably appropriate for use outside the state. Furthermore, it is stated that future editions will be even more so. Reliability coefficients for the 1929 edition vary from .55 for Section E and .56 for Section A, up to .92 for Sections GL and GF.

Percentile norms for the 1929 edition are:

Section	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
A	9	10	13	15	17	19	20
B-1	6	8	10	14	18	22	24
B-2	6	8	10	13	15	17	18
C	3	4	5	8	12	16	18
D	2	3	4	6	9	12	14
E	6	7	9	12	14	17	18
F	5	6	9	12	15	18	19
Total A–F	41	46	55	67	80	93	102
G—Latin	3	4	6	9	13	18	22
G—French	2	3	6	10	15	20	22
G—Agriculture	4	5	8	11	14	16	18

Bureau of Educational Research, University of North Carolina. Complete examination, 10¢ per copy, $1.25 per 25; single sections, 5¢ per copy, 50¢ per 25.

High School Achievement Test

W. W. D. Sones and D. P. Harry (1929)

Forms A, B

The four parts of this test are each divided into several sections. Part I, Language and Literature, has sixteen sections covering such phases of the subject as correct and faulty use of English word meaning, international authorship, familiar passages in literature, identification of grammatical and rhetorical forms, and so forth. In Part II, Mathematics, are eight sections dealing with mathematical processes, concepts, relationships, formulae, and so forth. Part III, Natural Science, includes ten sections on such phases of the work as classification, general principles, transformation of energy, scientific men and women, and so on. The last part, on the social studies, likewise includes ten sections which test knowledge of civic information, outstanding historical characters, international affairs, and so forth. The total number of elements is 140 in Part I, 80 in Part II, the same in Part III, and 115 in Part IV. Various forms are used, including multiple-answer, matching, single answer, and others. Both sections and elements within the sections are arranged in scalar order. The items included were determined largely by agreement among various committees and individuals as to subject matter.

Reliability data are given separately for the four parts. They are about as follows:

Part	r	$P.E._{meas.}$	$\frac{P.E._{meas.}}{M}$	$\frac{P.E._{meas.}}{\sigma}$
I	.94	4	.07	.16
II	.91	3	.12	.20
III	.90	3	.12	.21
IV	.93	4	.11	.18

No figures are given for the test as a whole, but evidently its reliability is higher than that of the single parts. Elaborate percentile norms for each part of the test and for each semester of high school are reported, and also according to the length of time each subject has been carried. Since they are stated to be only

tentative, they will not be reproduced at all completely, but merely the medians given.

Part	*Grade*							
	9B	*9A*	*10B*	*10A*	*11B*	*11A*	*12B*	*12A*
I	27	28	31	38	41	42	53	61
II	12	13	18	21	22	22	22	24
III	12	14	17	20	23	24	25	25
IV	18	21	22	23	27	27	35	32

Part	*Semesters carried*							
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
I	27	28	30	38	42	43	55	61
II	12	13	18	22	25	29	32	41
III	16	19	21	26	27	34	39	44
IV	17	20	19	25	28	30	36	—

World Book Company. Specimen set 30¢; $1.90 per 25.

BIBLIOGRAPHY

I. Agriculture

Smith, Z. M. "A Test of Animal Husbandry," *Thirteenth Annual Conference on Educational Measurements.* Bloomington: Bureau of Coöperative Research, Indiana University, 1926, p. 90–103.

II. Mechanics and Engineering

Bell, J. C. "Mechanical Aptitude and Intelligence," *Contributions to Education,* Vol. 1. Yonkers, New York: World Book Company, 1924, p. 270–82.

Chapman, J. C. and Chapman, Daisy. *Trade Tests.* New York: Henry Holt and Company, 1921. 435 p.

Moore, B. V. "Personnel Selection of Graduate Engineers," *Psychological Monographs,* Vol. 30, No. 5. Princeton, New Jersey: Psychological Review Company, 1921. 84 p.

Moore, B. V. "A Tested Method of Using Tests for Vocational Guidance," *School and Society,* 18:761–64, December 29, 1923.

Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction.* Yonkers, New York: World Book Company, 1927, p. 186–88.

Short, O. C. "Suggested Tests for Road Inspector," *Public Personnel Studies,* 3:252–66, September, 1925.

———. "Suggested Tests for Shift Engineman," *Public Personnel Studies,* 3:320–31, November, 1925.

Symonds, P. M. *Measurement in Secondary Education.* New York: The Macmillan Company, 1927, p. 383–88.

Toops, H. A. "Tests for Vocational Guidance of Children Thirteen to Sixteen," *Teachers College, Columbia University, Contributions to Education,* No. 136. New York: Bureau of Publications, Teachers College, Columbia University, 1923. 169 p.

————. "Trade Tests in Education," *Teachers College, Columbia University, Contributions to Education,* No. 115. New York: Bureau of Publications, Teachers College, Columbia University, 1921. 118 p.

"Information and Data Regarding Tests in the Short Answer Form," *Public Personnel Studies,* 6:219–20, October, 1928.

"Information and Data Regarding Tests in the Short Answer Form," *Public Personnel Studies,* 7:13–14, January, 1929.

"Objective Achievement Tests Constructed and Used in St. Louis," (*St. Louis*) *Public School Messenger,* 25:124–33, November 30, 1927.

"Suggested Tests for Automobile Mechanic," *Public Personnel Studies,* 4:232–46, August, 1926.

"Suggested Tests for Electrician," *Public Personnel Studies,* 4:106–16, March, 1926.

"Suggested Tests for Instrument Man," *Public Personnel Studies,* 5:39–44, February, 1927.

"Suggested Tests for Rodman," *Public Personnel Studies,* 6:67–71, March, 1928.

"Suggested Tests for Steam Fireman," *Public Personnel Studies,* 5:98–104, May, 1927.

III. General Survey Tests

Ruch, G. M. "A Mental-Educational Survey of 1550 Iowa High School Seniors," *University of Iowa Studies in Education,* Vol. 2, No. 5. Iowa City: University of Iowa, 1923. 29 p.

Stoddard, G. D. "Iowa Placement Examinations," *Journal of Engineering Education,* 16:158–63, October, 1925.

————. "Iowa Placement Examinations, Fall of 1925, Preliminary Report," *Journal of Engineering Education,* 16:475–83, March, 1926.

————. "Iowa Placement Examinations," *School and Society,* 24:212–16, August 14, 1926.

CHAPTER XV

GENERAL INTELLIGENCE

Introduction.—Although the term "general intelligence" is used as the title of this chapter and, therefore, to include the tests described herein, the writer recognizes that it is not an ideal expression for this purpose. The chief justification for its use is that it is the term most commonly employed and, therefore, most generally understood, in this connection. A number of other names for general intelligence tests, such as "mental tests," "mental alertness tests," "classification tests," and so forth, have been suggested, but none of these has been generally adopted.

The chief reason why general intelligence is not an entirely satisfactory expression to apply to tests of the sort described in this chapter is that there is no unanimity of opinion as to its meaning. Probably most persons have somewhat similar ideas called to mind by the term and think of somewhat the same traits or abilities as being referred to by it. When different psychologists and others have attempted to define it, however, there has been a marked lack of agreement. This is well illustrated by the statements from a number of our leading psychologists and workers with intelligence tests that appeared in a symposium published in 1921.[1] Of the definitions presented here and elsewhere the following are examples: "capacity to learn," "ability to solve problems," "ability to meet problematic situations," "ability to make adaptations to new conditions," "general mental adaptability to new problems and conditions of life," "intellect plus knowledge," "ability to profit by experience," and so forth. Although some critics have objected to the use of intelligence tests because of this fact, that there is no agreement as to just what they measure, this objection does not seem valid. As has fre-

[1] Thorndike, E. L., et al. "Intelligence and Its Measurement: A Symposium," *Journal of Educational Psychology*, 12:123–47, 195–216, 271–75; March, April, May, 1921.

quently been stated in rebuttal, no satisfactory definition of electricity has yet been given, yet we do not hesitate to employ it for many purposes in everyday life and to derive a great deal of benefit from its use. Similarly the mere fact that we cannot define exactly what they measure does not prevent us from employing the results yielded by intelligence tests in ways that serve to improve the educational process.

Both before general intelligence tests were devised and concurrent with them numerous tests of various mental abilities and capacities have been employed. Sometimes the general term "psychological tests" is applied to these, although frequently this includes intelligence tests also or is applied to them alone. The tests referred to include tests of memory, learning, imagination, association, ability to observe and report, general information, and so forth. Such tests are not dealt with in this chapter except as they form parts of so-called general intelligence tests. The chief reason for this is that psychologists and research workers rather than teachers or others engaged directly in public-school education are primarily interested in them. The tests that are considered are grouped in two divisions, individual and group. None of the comparatively few non-verbal tests, that is, tests in which knowledge of language is not required, are included. Such tests have their place, but are rarely profitably employed in testing high-school pupils, since such pupils are almost always able to read and otherwise employ language well enough that verbal tests may be satisfactorily given.

The tests to be described in this chapter have been selected from a larger number actually available than is true in the case of any of the high-school subjects. Approximately two hundred general intelligence tests have been constructed, of which around one-half can probably be secured in sufficient quantities for administration to pupils. Of course, not all of these are suitable for the high-school level, but yet the number appropriate therefor is rather large. Because of this condition, that so many tests already exist, and because of the further fact that the best tests now available apparently represent rather well the best knowledge as to how they should be made, there is at present and has been during the past few years much less activity in the construction of general intelligence tests than in connection with many of the

high-school subjects. Therefore, as will be seen, the average date at which the tests included in this chapter appeared is several years earlier than in almost any other chapter of this book.

One point that should be emphasized in connection with intelligence tests as contrasted with achievement tests is the much greater importance of following directions exactly and of giving no help not specified therein. It is assumed that the directions on achievement tests make clear to pupils just how they are to respond so that such tests measure their knowledge or ability in the subjects dealt with and not their ability to understand directions. In intelligence tests, on the other hand, the directions are frequently thought of as constituting a part of the test, and the ability to understand and follow them should, therefore, play a part in determining the score. This is not uniformly true, since in the case of some tests the authors had no such intention, whereas at the other extreme there are some general intelligence tests that are almost entirely tests of ability to follow directions. Since the person giving the test is frequently not aware of just how much of this element is supposed to enter into determining the scores, it is, as stated above, very important that he adhere exactly to the directions for administering tests.

In connection with tests in the school subjects, it is, and seemingly always has been, a tacit assumption that teachers and usually pupils also should know just what scores are made. In the case of intelligence tests, however, the situation is decidedly different. Some of the leading workers in this field recommend that the scores made by pupils thereon be regarded as highly confidential, and in ordinary cases be known only by the chief administrative or supervisory officials directly in charge of the pupils, the person or persons who have actually given the tests, and personnel or guidance workers. Others advise that intelligence-test scores ordinarily be revealed to teachers, arguing that the better teachers know their pupils the more efficient they can be. The chief argument advanced against this practice is that teachers are too liable to be prejudiced, especially against inferior pupils. Practically all are agreed that except in unusual circumstances it is unwise to reveal to pupils and generally also to their parents, their intelligence test scores. The writer's observation and experience have led him to feel that in many cases at least

it is wise to give teachers access to the intelligence ratings of their pupils and that practically all teachers worthy of the name will, on the whole, profit by having this information. Teachers who will be prejudiced against inferior pupils because of their inferiority, or who will unduly favor superior pupils because of their superiority, will in most cases exhibit the same favorable and unfavorable prejudices on other grounds, and thus the situation will be no worse if they are informed of their pupils' intelligence ratings. On the other hand, the writer believes that the vast majority of conscientious teachers will profit by knowing how their pupils differ in intelligence, will aid them more efficiently because they understand them better, will do more to stimulate the bright and exhibit more patience with the dull, and on the whole adapt the work offered in their instruction to the capacities of their pupils much better than if they do not have this information.

With regard to revealing the results of such tests to pupils and parents, the writer is in accord with general opinion. He wishes to point out, however, that in some exceptional cases it is desirable to do so. Some pupils who think they cannot master certain subjects can be stimulated to do so if they are informed that their intelligence is really average or above, and others who are mentally lazy can be stimulated to better work by being given this same fact in rather pointed fashion. Likewise pupils below average sometimes become discouraged because they cannot keep up with their more gifted classmates and by letting them know that they are perhaps doing even better with regard to their ability, much of this discouragement may be removed. The same is true with regard to parents. Sometimes if they realize that their children are above average they will better stimulate them to work up to their capacity; also if they realize they are below average they will be more sympathetic with their shortcomings and not, as is sometimes the case, blame and punish them severely when they are doing all that can reasonably be expected of them.

I. Individual Intelligence Scales

As has already been stated in Chapter II, the first general-intelligence tests were individual in their nature, that is, must be administered singly and not to groups. The first scale of this sort,

in the modern sense of the term, was the well-known Binet-Simon Intelligence Scale, which first appeared in 1905 [2] and was revised in 1908 [3] and 1911.[4] Practically all individual intelligence scales involving verbal elements now employed either in this or other countries are revisions or extensions of the work of Binet and Simon, or in some way have their origins therein. In some cases the lines of development, although started at the same point, have been so different that the results can hardly be recognized as springing from the same source unless one knows their history. On the other hand, in most of those that are receiving wide use in this country at the present time many common elements are easily recognizable.

From one standpoint, at least, individual intelligence scales are always better than group intelligence tests. They yield better measures of what it is desired to measure, being usually more reliable and giving more detailed and diagnostic information concerning the mentality of those to whom they are administered. The chief reason why they are not commonly employed is that they are too time consuming. Practically none of them can be given satisfactorily to high-school pupils in less than forty minutes, and some require considerably more than this amount of time. A second reason is that to give and score such tests correctly requires the services of testers considerably better trained and more expert than is necessary with group tests. Some psychologists and others have overemphasized this fact, stating that no one should be allowed to administer individual intelligence tests who has not had a complete course dealing with them, and a considerable amount of supervised practice. This, of course, is desirable, but it is possible for an individual who has a good fundamental knowledge of educational psychology and possesses average or better ability, to train himself by studying the proper materials and practicing, if possible, on children who have already been tested by experts, so that he can administer individual tests well enough

[2] Binet, A. et Simon, T. "Methodes Nouvelles pour le Diagnostic du Niveau Intellectuel des Anormaux," *L'Année Psychologique*, 11:191–244, 1905.

[3] ———— "Le Développement de l'Intelligence chez les Enfants," *L'Année Psychologique*, 14: 1–90, 1908.

[4] Binet, A. "Nouvelles Recherches sur la Mesure du Niveau Intellectuel chez les Enfants d'École," *L'Année Psychologique*, 17:145–201, 1911.

to justify his doing so. Most high-school teachers, however, are not called upon to do this work, so although a general knowledge of such tests is desirable, they should not feel called upon to attempt to learn how to administer them.

It is commonly recommended, and the writer agrees therewith, that in the ordinary school situation individual intelligence tests be given to a relatively small proportion of pupils. Those at the lowest and highest extremes of intelligence, those who are unusual or problem cases, those for whom the results from two or more group intelligence tests disagree widely, or in whose cases there is such a disagreement between the results of intelligence tests and teachers' estimates of intelligence, and perhaps others, are the ones to whom individual tests should be administered.

Stanford Revision of the Binet-Simon Intelligence Tests

L. M. Terman et al. (1916) [5]

Tests for Years III, IV, V, VI, VII, VIII, IX, X, XII, XIV Average Adult, Superior Adult [6]

For each year named above there is a group of tests, usually six, but in one case eight, with from one to three alternative tests in all but two cases. The tests in the various groups are as follows:

Year III (six tests, two months each)
1. Identify parts of body.
2. Name familiar objects.
3. Enumerate objects in picture.
4. Give sex.
5. Give last name.
6. Repeat six or seven syllables.

Alt. Repeat three digits.

Year IV (six tests, two months each)
1. Compare lines.
2. Discriminate forms.

[5] Although Terman and his co-workers published a tentative revision and extension of the Binet-Simon Scale in 1912, the form in which it has been widely used was not made available until 1916.

[6] The average adult test is equated to a mental age of sixteen and one-half years and the superior adult to one of nineteen and one-half years.

3. Count four cents.
4. Copy square with pencil.
5. Respond to easy questions.
6. Repeat four digits.

Alt. Repeat twelve or thirteen syllables.

Year V (six tests, two months each)

1. Compare weights.
2. Name colors.
3. Make esthetic comparison of pictures.
4. Define names of common objects.
5. Form rectangle from two triangles.
6. Execute three easy commissions.

Alt. Give age.

Year VI (six tests, two months each)

1. Distinguish right and left.
2. Name omissions from pictures.
3. Count thirteen cents.
4. Respond to questions.
5. Name different coins.
6. Repeat sixteen to eighteen syllables.

Alt. State whether it is morning or afternoon.

Year VII (six tests, two months each)

1. State number of fingers.
2. Describe pictures.
3. Repeat five digits.
4. Tie bow-knot.
5. Give differences between objects named.
6. Copy diamond with pen.

Alt. 1. Name days of week in order.

Alt. 2. Repeat three digits backward.

Year VIII (six tests, two months each)

1. Draw plan for finding ball in field.
2. Count from twenty down to one.
3. Respond to questions.
4. Give similarities between two things named.
5. Give definitions superior to use.
6. Indicate vocabulary by definition.

Alt. 1. Name different coins.

Alt. 2. Write from dictation with pen.

Year IX (six tests, two months each)

1. Give date.
2. Arrange weights in order.
3. Make change.
4. Repeat four digits backward.
5. Make up sentences containing given words.
6. Give rhymes for given words.

Alt. 1. Name months in order.

Alt. 2. State total value of several stamps.

Year X (six tests, two months each)
1. Indicate vocabulary by definition.
2. Point out absurdities in statements.
3. Copy designs.
4. Read selection and then tell in own words what has been read.
5. Respond to questions.
6. Name words.

Alt. 1. Repeat six digits.
Alt. 2. Repeat twenty to twenty-four syllables.
Alt. 3. Solve form-board puzzle.

Year XII (eight tests, three months each)
1. Indicate vocabulary by definition.
2. Define abstract words.
3. Draw plan for finding ball in field.
4. Rearrange dissected sentences properly.
5. Interpret fables.
6. Repeat five digits backward.
7. Interpret pictures.
8. Give similarities of groups of three things each.

Year XIV (six tests, four months each)
1. Indicate vocabulary by definition.
2. Determine rule by induction.
3. Distinguish between president and king.
4. Respond to problem questions.
5. Solve reasoning problems in arithmetic.
6. Visualize and state results from reversing hands of clock.

Alt. 1. Repeat seven digits.

Average Adult (six tests, five months each)
1. Indicate vocabulary by definition.
2. Interpret fables.
3. State differences between abstract terms.
4. Solve problem concerning enclosed boxes.
5. Repeat six digits backward.
6. Learn and employ code for writing.

Alt 1. Repeat twenty-eight syllables.
Alt. 2. Comprehend physical relations.

Superior Adult (six tests, six months each)
1. Indicate vocabulary by definition.
2. Visualize results from folding and cutting paper.
3. Repeat eight digits.
4. Repeat thought of difficult passage.
5. Repeat seven digits backward.
6. Solve puzzle problem.

It will be noticed that in the list of tests given above the same or very similar ones are often repeated at different ages. In such cases the test at the higher age involves more difficult material or

else better responses. For example, Test No. 5 at Year IV requires satisfactory answers to the questions "What must you do when you are sleepy? When you are cold? When you are hungry?" whereas Test No. 4 at Year VI inquires what is to be done "if it is raining when you start to school, if you find that your house is on fire," and "if you are going some place and miss your car?" To give another example, the vocabulary test that appears at ages VIII, X, XII, XIV, Average and Superior Adult, is considered as passed if twenty words are satisfactorily defined at age VIII, thirty at age X, and so on up to seventy-five at the Superior Adult level.

The ninety tests in the Stanford Revision include the fifty-four that were in the Binet-Simon 1911 Revision, some of them, however, being considerably modified, and thirty-six new ones, most of which were constructed by Terman and his associates. The material was tried out on about seventeen hundred normal children, two hundred defective and superior children, and four hundred adults. The age placement of the tests was according to the principle that average children of a given chronological age should score exactly at the same mental age.

In the case of a number of the tests there are time limits within which the pupils' responses must be given to be credited. In others there are no definite time limits. In some cases the responses to a question or exercise must be entirely correct, whereas in others a certain approximation to correctness is sufficient for credit to be given. The total time required to give the tests varies considerably according to the speed at which the examiner works. For most high-school pupils a satisfactory test cannot be completed in less than forty or forty-five minutes, and for very few is it necessary to use more than one hour.

Although the validity of the Stanford Revision has been attacked on several counts, the chief one of which is probably that it includes too much verbal material, it is generally considered the best, and certainly is the most widely used, individual intelligence test in the English language. It is true that many of the tests that compose it deal with verbal material and are more or less closely connected with achievement in school subjects, but it is no more true of the Stanford than of practically all verbal intelligence tests. The position it holds among tests is perhaps best

attested by the fact that it has very commonly been taken as the criterion measure or standard of comparison for other individual and group intelligence tests.

Many data have been published as to the reliability of the Stanford Revision both in terms of mental ages and of I. Q.'s. An average of a number of these investigations yields about the following data for mental ages:

$$r = .95,\ \text{P.E.}_{\text{meas.}} = 6 \text{ months},\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .05,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .16.$$

For I.Q.'s they are practically the same, except that the probable error of measurement is only about five points. These figures are for re-tests within a comparatively short time by the same examiner or by different examiners whose procedure is very similar. For longer periods of time running into months or years and for examiners who differ more in their methods, the reliability is somewhat less, although the decrease is not as great as might be supposed.

Since each test counts a definite number of months of mental age, there are no norms in the ordinary sense of the term. Pupils are given credit for the mental age at the highest year level at which they pass all tests and for the additional amount represented by the months credited to all tests above that year passed. For example, if a pupil passes all the twelve-year-old tests, four of those at fourteen, and two at average adult level, his mental age is fourteen years, two months, computed as follows: twelve years (since all tests at that level were passed), plus sixteen months for the four fourteen-year-old tests which count four months each, plus ten months for the two average adult tests, which count five months each.

C. H. Stoelting Company. Complete set of material $8.55; complete instructions $2.95; condensed guide $1.30; record booklets $2.50 per 25; abbreviated filing record blanks $1.30 per 25.

Also Houghton Mifflin Company. Test material $1.00; condensed guide $1.00; record booklets $2.00 per 25; abbreviated file record cards $1.00 per 25.

References: Terman, L. M. and Childs, H. G. "A Tentative Revision and Extension of the Binet-Simon Measuring Scale of Intelligence," *Journal*

of Educational Psychology, 3:61–74, 133–43, 198–208, 277–89; February, March, April, May, 1912.

Terman, L. M. *The Measurement of Intelligence*. Boston: Houghton, Mifflin Company, 1916. 362 p.

Terman, L. M., et al. *The Stanford Revision and Extension of the Binet-Simon Scale for Measuring Intelligence*. Baltimore: Warwick and York, 1917. 179 p.

Terman, L. M. *The Intelligence of School Children*. Boston: Houghton Mifflin Company, 1919. 317 p.

Revision of the Binet-Simon Tests

J. P. Herring (1922)

Form A; Groups A, B, C, D, E

This is a series of thirty-eight tests very similar to those included in the Stanford Revision, but in many cases more inclusive, that is, one test in this frequently contains several parts corresponding to several similar tests appearing at various levels of difficulty in the Stanford. Group A includes the first four tests of the series, Group B the first thirteen, C the first twenty-two, D the first thirty-one, and E the complete series. The purpose of having these groups is that the examiner is to give the longest group for which time is available, preferably, of course, Group E. Apart from the combination of similar tests into one, the chief difference between this and the Stanford is that only printed material, practically all of which is contained in a single booklet, is required, so that no such objects as weights, knife, key, coins, and so forth, are employed. The time required to give the complete series is about the same as for the Stanford, perhaps a little less, and that for each of the shorter groups approximately proportional to the number of tests they include. Group A can be given in from five to ten minutes, and so on up to perhaps forty or fifty for E. The scoring system used differs from that of the Stanford in that point scores are determined and transmuted into equivalent mental ages.

Apparently the validity of this test is equal to that of the Stanford. Indeed, its author and others frequently refer to it as a duplicate form thereof rather than as a separate test. This opinion is supported by the high correlation between the two. It is not, however, as valid at its lowest end, but this does not affect its use in high school. Herring gives reliability data based upon the com-

parison of scores on his revision with those on the Stanford. He claims the following:

$$r = .99,\ \text{P.E.}_{\text{meas.}} = 3,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .03,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .07.$$

These data refer to I.Q.'s. Another investigator has reported slightly lower agreement between this and the Stanford as follows:

$$r = .97,\ \text{P.E.}_{\text{meas.}} = 4,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .04,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .12.$$

The shorter groups correlate highly with the longer; Group A .95 with E, B .97, C and D slightly more.

It is somewhat doubtful whether it is better to employ the Herring Revision rather than to administer the Stanford Revision a second time to pupils who have already been tested with the latter. There appears to be little increase of scores due to practice effect when the Stanford Revision is given to pupils previously tested who do not expect it again, especially if a few months or even weeks have elapsed. On the other hand, since the Herring Revision correlates so highly with the Stanford, it is perhaps just as well to employ it the second time. In doing so, however, one should recognize that because of the great similarity between the two, so doing will not entirely obviate the danger of practice effect. It is slightly easier to give than the Stanford because of the grouping of similar tests, and the fact that miscellaneous material is not needed.

The numbers of points on each of the five groups equivalent to the mental ages that include most high-school pupils are as follows:

Group	*Mental age*						
	12	*13*	*14*	*15*	*16*	*17*	*18*
A	25	27	29	31	33	35	37
B	60	64	68	72	76	80	84
C	91	98	106	114	122	130	—
D	111	120	129	137	146	155	164
E	138	148	158	167	177	187	196

World Book Company. Manual $1.00; individual record cards $1.00 per 25.

References: Herring, J. P. "Herring Revision of the Binet-Simon Tests," *Journal of Educational Psychology*, 15:172–79, March, 1924.

————. "Reliability of the Stanford and the Herring Revision of the Binet-Simon Tests," *Journal of Educational Psychology*, 15:217–223, April, 1924.

————. "Avery's Comparison of the Stanford and Herring Revisions," *Journal of Educational Psychology*, 15:383–88, September, 1924.

Avery, G. T. "Comparison of Stanford and Herring Binet Revisions Given to First Grade Children," *Journal of Educational Psychology*, 15:224–28, April, 1924.

Wilner, C. F. "A Comparative Study of the Stanford and the Herring Revisions of the Binet-Simon Tests," *Journal of Educational Psychology*, 15:520–29, November, 1924.

Revision and Extension of the Binet-Simon Scale

F. Kuhlmann (1922)

Ages 3, 6, 12, 18 months; II, III, IV, V, VI, VII, VIII, IX, X, XI, XII, XIII–XV years

This is a revision of an earlier set of tests by the same author. Although in many ways similar to the Stanford and Herring Revisions, especially to the first in the arrangement of tests at age levels, it differs from both somewhat more than either of them does from the other. There are five tests at each of the ages given in months and eight at each of those in years.

The time required is slightly longer than for the Stanford. This series of tests will not measure as high mental ability as will either of the others previously described in this section. It is, therefore, not suitable for use through the high school, but only in the first year or two, or with decidedly inferior pupils. The plan of scoring is similar to that used with the Stanford in that each test passed counts for a certain number of months of mental age. At three and six months each counts six-tenths months, at twelve and eighteen months and two years, 1.2 months, and from three years up, 1.5 months. Extensive data on reliability have not been published, but it appears that this series is about equal to the Stanford in that respect. On the whole, there appears to be no sufficient reasons to recommend the use of Kuhlmann's Revision in preference to the Stanford and the Herring in the high school, but it is among the three or four best individual intelligence tests available.

Warwick and York. Complete outfit and 50 record sheets $7.00.

References: Kuhlmann, F. "A Revision of the Binet-Simon System for Measuring the Intelligence of Children," *Journal of Psycho-Asthenics, Monograph Supplement*, Vol. 1, No. 1. September, 1912. 41 p.

———. *A Handbook of Mental Tests.* A Further Revision and Extension of the Binet-Simon Scale. Baltimore: Warwick and York, 1922. 208 p.

Goodenough, F. L. *The Kuhlmann-Binet Tests for Children of Preschool Age.* Minneapolis: University of Minnesota, 1928. 146 p.

II. Group Intelligence Tests

Because of the greater economy of time in administration by far the greater number of intelligence tests actually employed in school are group tests. Most of them can be given to groups of pupils of almost any size for which there are seating accommodations in little if any more than the time required for one individual test. Moreover, less training and skill is required to give and score group than individual tests. Because of these facts, group tests have almost entirely superseded individual tests for the measurement of large numbers of children and to a considerable extent when individuals are being dealt with. Frequently instead of giving an individual test, several group tests are given. The purpose of the examiner determines which is preferable. If this is chiefly to determine a mental age, intelligence quotient, or other similar measure, regardless of variations in the different phases of mental ability, an average of several group tests is probably better than the result from one individual test. If, however, the chief purpose is to diagnose and analyze a pupils' mental ability, an individual test is more helpful.

A common method of validating group intelligence tests has been to compare them with an individual scale, usually the Stanford Revision. Very few group tests yield correlations much higher than .75 with this criterion. The same is true of intercorrelations among group tests. These usually range from about .75 down to .50. Also teachers' estimates of intelligence have frequently been used as a basis of comparison. The resulting correlations are usually somewhat lower than those just quoted.

Out of the large number of group intelligence scales, five suitable for use during most or all of the high-school period and two others that fit the end of this period have been selected for inclusion in this chapter.

Group Intelligence Scale
A. S. Otis (1918)
Advanced Examination; Forms A, B

This was the first group intelligence scale primarily intended for use in the secondary school to appear. Also it has probably received the widest use of any in Grades VII to XII. The scale consists of ten subtests as follows: following directions, opposites, disarranged sentences, proverbs, arithmetic (reasoning problems), geometrical figures, analogies, similarities, narrative completion, and memory. These include 230 elements in single-answer, multiple-answer, true-false, and other forms. The total time that the pupils actually work is about forty-five minutes, but the directions to the various subtests are so long that the test requires over an hour to give.

Although, as stated above, this was the first test of this sort to be made available, it still ranks among the best, and few experts in the field would rank it lower than third or fourth place, some even higher than that. Coefficients of reliability for the different half grades and grades are reported ranging from .37 to .97, averaging about .84. Another report gives a coefficient of .90 for two or three grades combined. For all grades combined the following figures appear to be the best available.

$$r = .97,\ \text{P.E.}_{\text{meas.}} = 7,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .06,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .12.$$

A correlation of .75 with the Stanford Revision, an average of .71 with thirteen other tests, and of .92 with the composite of all have been found. Otis gives the following equivalent mental ages according to the Stanford Revision for point scores on his test.

Point score	*80*	*90*	*101*	*112*	*125*	*140*	*160*	*183*
Mental age	12	13	14	15	16	17	18	19

The median scores actually made by individuals at the various ages are somewhat different. These are as follows:[1]

[1] The norms for all of the tests described in this section are based on many thousands of cases.

Age	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*
Point score	80	90	100	110	120	127	130	130

Also grade percentile norms are given.

Grade	*Percentile*				
	10	*25*	*50*	*75*	*90*
XII	120	136	154	169	183
XI	113	131	149	167	184
X	106	120	130	153	166
IX	92	108	125	142	157
VIII	80	95	112	128	144
VII	65	80	97	113	128

World Book Company. Specimen set 50¢; $1.25 per 25; manual 30¢.

References: Otis, A. S. "An Absolute Point Scale for the Group Measurement of Intelligence," *Journal of Educational Psychology*, 9:239–61, 333–48; May, June, 1918.

Colvin, S. S. "Some Recent Results Obtained from the Otis Group Intelligence Scale," *Journal of Educational Research*, 3:1–12, January, 1921.

Group Test of Mental Ability
L. M. Terman (1920)
Forms A, B

This test, which is intended for Grades VII to XII, is more often rated as the best for these grades than is any other group test. It is perhaps slightly too easy for the brightest pupils in high school, but this shortcoming is not serious. It consists of ten subtests as follows: information, best-answer, word meaning, logical selection, arithmetic (reasoning), sentence meaning, analogies, mixed sentences, classification, and number series. A total of 185 items make up these subtests. The proportion of alternative exercises is somewhat greater than in most tests, although not so great as that of multiple-answer ones. The final forms represent a rather careful tryout of almost twice as many items as were finally used. The actual working time is twenty-seven minutes, and the directions short enough that the whole may be given in from thirty-five

to forty, thus being convenient for use in connection with high-school periods of the ordinary length.

Many reliability data have been reported, some of which are not at all in agreement with others. The following appears to be an average statement, however:

$$r = .90,\ \text{P.E.}_{\text{meas.}} = 4,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .04,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .21.$$

Scores correlate at least .75, perhaps slightly better, with those on the Stanford Revision. An average correlation of .75 with thirteen other tests and of .92 with the total of all has been found. Mental age norms are as follows:

Point score	*50*	*72*	*93*	*113*	*135*	*157*	*178*	*200*
Mental age	12	13	14	15	16	17	18	19

Percentile norms are given as follows:

Grade	*Percentile*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
XII	194	185	169	147	122	100	86
XI	189	180	163	138	112	90	77
X	177	166	147	122	98	79	67
IX	164	151	128	104	81	63	53
VIII	148	135	112	89	69	52	43
VII	122	109	88	68	51	38	31

World Book Company. Specimen set 15¢; $1.20 per 25.

Mental Ability Test
W. S. Miller (1921)
Forms A, B

This is a considerably shorter test than either of the two previously described. It consists of only three subtests, each of which calls for forty responses. Actual working time is nineteen minutes, and the total time required less than thirty. The fact that it is shorter naturally makes it somewhat less valid and reliable than many other tests, but in proportion to its length it is not inferior

in these qualities. Indeed, it is frequently recommended as the best distinctly short test for high-school use. Its reliability appears to be stated approximately by the following figures:

$$r = .91,\ \text{P.E.}_{\text{meas.}} = 7,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .11,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .20.$$

Its correlation with the average of five intelligence tests including itself has been found to be .90, and its average correlation with the others .76. Form B has been found to be slightly more than a year more difficult than A. Mental age norms are not provided as such, but apparently are roughly as follows:

Point score	*25*	*36*	*47*	*55*	*64*	*71*	*78*	*90*
Mental age	12	13	14	15	16	17	18	19

Percentile norms are as follows:

Grade	*Percentile*				
	10	*25*	*50*	*75*	*90*
XIII	60	72	85	95	103
XII	50	61	74	85	93
XI	44	56	69	80	89
X	39	50	62	73	84
IX	29	40	53	66	77
VIII	25	35	46	59	68
VII	17	25	35	46	58

World Book Company. Specimen set 25¢; 80¢ per 25; manual 15¢.

Self Administering Tests of Mental Ability

A. S. Otis (1922)

Higher Examination; Forms A, B, C

This is an omnibus test calling for seventy-five responses to multiple-answer and single-answer exercises. It is intended to be given with a minimum of directions by the examiner, who need do nothing more than tell the pupils to fill in the blanks for name and so forth, read the directions to themselves, start when the signal is

given, and stop at the proper time. Provision is made for giving it with either twenty or thirty minutes of working time. The latter is preferable in the lower years of high school and probably in the upper also, although a few of the brightest juniors and seniors will complete it in less than thirty minutes.

This test is a product of very careful work, and despite the fact that its length is not great, ranks high with respect to both validity and reliability. One reason for this is that the opportunity for differences in administration due to errors on the part of teachers or others giving the test is much less than with most group intelligence tests.

$$r = .92,\ \text{P.E.}_{\text{meas.}} = 2.6,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .06,\ \frac{\text{P.E.}_{\text{meas}}}{\sigma} = .19.$$

A correlation of .89 with the same author's Advanced Examination has been obtained. Apparently Form B is about four points more difficult than A. In the case of this test, as of his other one, Otis gives no mental age norms above eighteen. Up to this they are as follows:

Point score	*23*	*28*	*32*	*36*	*39*	*41*	*42*
Mental age	12	13	14	15	16	17	18

Grade norms are also given:

Grade	*VIII*	*IX*	*X*	*XI*	*XII*
Norm	36	41	45	48	50

In both cases these are for thirty minute time limits. The manual gives a table by which twenty-minute scores may be made equivalent to those from thirty-minute testing. Although provision is made for determining intelligence quotients, Otis also provides for securing indices of brightness from the point scores on this test.

World Book Company. Specimen set 30¢; 80¢ per 25.

Senior Tests
S. L. and L. C. Pressey (1922)
Classification, Verifying Tests

The two tests named above are for all practical purposes two duplicate forms of the same test. Each consists of ninety-six multiple-answer elements, most of which deal with what may be called general information, including vocabulary, arithmetic, history, matters of everyday life, and so forth. Sixteen minutes are allowed, which makes these the shortest of the group tests described in this chapter. Because of the similar form of all the exercises, it is among the easiest to give, one set of directions covering the whole.

Data on reliability have not been reported, but there appears to be no reason why this should not be almost as reliable as the Miller. Age norms are as follows:

Point score	*8*	*16*	*27*	*40*	*50*	*60*	*69*
Mental age	12	13	14	15	16	17	18

Grade norms for the end of the year are:

Grade	*VII*	*VIII*	*IX*	*X*	*XI*	*XII*
Norm	24	38	47	56	66	79

Public School Publishing Company. Sample set 10¢; $1.25 per 100.

Reference: Pressey, S. L. and Pressey, L. C. "A Revision of the Pressey Primer and Cross-Out Scales," *Journal of Educational Research*, 6:178–79, September, 1922.

Brown University Psychological Examination
S. S. Colvin (1919)
Fore-exercise D; Exercises E, F

This is perhaps the best test for high-school seniors and college freshmen among those that do not much exceed an hour in length. The fore-exercise contains four subtests which illustrate the types of those in the two regular exercises. The first of the latter con-

tains four and the second five parts. These include completion language exercises, definitions, opposites, and other similar material, all of which is highly verbal except one subtest that deals with arithmetic. The total time is seventy minutes, but the fore-exercise and E may be given at one time and F at another so that it will fit the ordinary high-school period. A coefficient of reliability of .84 is reported. For this as for practically all intelligence tests intended for college use it is not customary to use mental-age scores, intelligence quotients, or other similar derived measures.

J. B. Lippincott Company. $1.50 per 25; manual 15¢.

Reference: Colvin, S. S. "The Validity of Psychological Tests for College Entrance," *Educational Review*, 60:7–17, June, 1920.

Intelligence Examinations for High-School Graduates

E. L. Thorndike (1919)

This is the longest and probably the best test that has received any considerable amount of use for measuring high-school seniors and college freshmen. Three issues with different content are published each year. Each consists of five parts and calls for several hundred responses in all. Many types of exercises are employed. They include true-false statements, following directions, arithmetical computation, completion of number series, use of symbols, definitions, and so forth. The total working time is almost three hours, and at least thirty minutes more are required for directions and other details of administration. In this as in the Brown, there is a ten-minute practice test. It is recommended that if a shorter test is desired Part I, which contains 136 items and can be given in about a class period, be employed.

$$r = .85,\ \text{P.E.}_{\text{meas.}} = 3,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .05,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .26.$$

In view of the length of the test, these figures do not seem high, but rather lower than one would expect. Intercorrelations among the various parts range from .75 to .90.

Bureau of Publications. Specimen set 50¢; complete sets, class order only, current issues 75¢ per copy; back issues 50¢ per copy.

References: Thorndike, E. L. "Intelligence Examinations for College Entrance," *Journal of Educational Research*, 1:329–37, May, 1920.
Root, W. T. "The Freshman: Thorndike College Entrance Tests, First Semester Grades, Binet Tests," *Journal of Applied Psychology*, 7:77–92, March, 1923.
Wood, B. D. "Description, Validity, and Analysis of the Thorndike Examination," *Measurement in Higher Education*. Yonkers, New York: World Book Company, 1923, Chapters III, IV, and V.

BIBLIOGRAPHY

General

Bronner, A. F., et al. *A Manual of Individual Mental Tests and Testing*. Boston: Little Brown and Company, 1927. 287 p.
Colvin, S. S. (Chairman). "Intelligence Tests and Their Use," *Twenty-First Yearbook of the National Society for the Study of Education*, Parts I and II. Bloomington, Illinois: Public School Publishing Company, 1922. 288 p.
Dearborn, W. F. *Intelligence Tests, Their Significance for School and Society*. Boston: Houghton Mifflin Company, 1928. 336 p.
Dewey, Evelyn, Child, Emily, and Ruml, Beardsley. *Methods and Results of Testing School Children*. New York: E. P. Dutton and Company, 1920. 176 p.
Franz, S. I. *Handbook of Mental Examination Methods*, Second Edition, Revised and Enlarged. New York: The Macmillan Company, 1919. 193 p.
Freeman, F. N. *Mental Tests, Their History, Principles and Applications*. Boston: Houghton Mifflin Company, 1926. 503 p.
Gregory, C. A. "The Measurement of Intelligence," *Fundamentals of Educational Measurement*. New York: D. Appleton and Company, 1922, Chapters III and IV.
Herring, J. P. "Verbal and Abstract Elements in Intelligence Examinations," *Herring Revision of the Binet-Simon Tests, and Verbal and Abstract Elements in Intelligence Examinations*, Part II. Yonkers, New York: World Book Company, 1924, p. 18–71.
Hines, H. C. *Measuring Intelligence*. Boston: Houghton Mifflin Company, 1923. 146 p.
Levine, A. J. and Marks, Louis. *Testing Intelligence and Achievement*. New York: The Macmillan Company, 1928, Chapters I, III, IV, and V.
Peterson, Joseph. *Early Conceptions and Tests of Intelligence*. Yonkers, New York: World Book Company, 1925. 320 p.
Pintner, Rudolf. *Intelligence Testing, Methods and Results*. New York: Henry Holt and Company, 1923. 406 p.
Seashore, C. E., et al. "Mentality Tests: A Symposium," *Journal of Educational Psychology*, 7:229–40, 278–86, 348–60; April, May, June, 1916.
Stern, William. (Translated from the German by G. M. Whipple). *The Psychological Methods of Testing Intelligence*. Baltimore: Warwick and York, 1914. 160 p.

Thorndike, E. L., et al. "Intelligence and Its Measurement: A Symposium," *Journal of Educational Psychology*, 12:123–47, 195–216, 271–75; March, April, May, 1921.

———. *The Measurement of Intelligence*, New York: Bureau of Publications, Teachers College, Columbia University, 1926. 616 p.

Wells, F. L. *Mental Tests in Clinical Practice*. Yonkers, New York: World Book Company, 1927. 315 p.

Whipple, G. M. *Manual of Mental and Physical Tests*, Parts I and II, Second Edition, Revised and Enlarged. Baltimore: Warwick and York, 1915. Part I, 365 p.; Part II, 336 p.

Wylie, A. T. "A Brief History of Mental Tests," *Teachers College Record*, 23:19–33, January, 1922.

Yerkes, R. M., et al. "Psychological Examining in the United States Army," *Memoirs of the National Academy of Sciences*, Vol. 15. Washington: Government Printing Office, 1921. 890 p.

Yoakum, C. S. and Yerkes, R. M. (Compiled and Edited by). *Army Mental Tests*. New York: Henry Holt and Company, 1920. 303 p.

Young, Kimball. "The History of Mental Testing," *Pedagogical Seminary*, 31:1–48, March, 1924.

I. Individual Intelligence Scales

Arthur, Grace. "Clinical Manual," *A Point Scale of Performance Tests*, Vol. 1. New York: The Commonwealth Fund, Division of Publications, 1930. 82 p.

Burt, Cyril. *Handbook of Tests for Use in Schools*. London, England: P. S. King and Son, 1923. 106 p.

———. *Mental and Scholastic Tests*. London, England: P. S. King and Son, 1921. 432 p.

Goddard, H. H. "The Binet and Simon Tests of Intellectual Capacity," (*Vineland, New Jersey*) *Training School Bulletin*, 5:3–9, December, 1908.

———. "Four Hundred Feeble-Minded Children Classified by the Binet Method," *Pedagogical Seminary*, 17:387–97, September, 1910.

———. "A Revision of the Binet Scale," (*Vineland, New Jersey*) *Training School Bulletin*, 8:56–62, June, 1911.

———. "Two Thousand Children Measured by the Binet Measuring Scale of Intelligence," *Pedagogical Seminary*, 18:232–59, June, 1911.

Kohs, S. C. *Intelligence Measurement*. New York: The Macmillan Company, 1923. 312 p. (Revised 1927.)

Kubo, Y. "The Revised and Extended Binet-Simon Tests, Applied to the Japanese Children," *Pedagogical Seminary*, 29:187–94, June, 1922.

Melville, N. J. *Testing Juvenile Mentality, Second Enlarged Edition*. Philadelphia: J. B. Lippincott Company, 1920. 162 p.

Phillips, G. E. *The Measurement of General Ability, An Australian Revision and Extension of the Binet-Simon Scale*. Sydney, New South Wales: Teachers' College Press, 1924.

Pintner, Rudolf and Paterson, D. G. *A Scale of Performance Tests*. New York: D. Appleton and Company, 1917. 218 p.

Squires, P. C. *A Universal Scale of Individual Performance Tests Examination Manual.* Princeton, New Jersey: Princeton University Press, 1926. 158 p.

Trabue, M. R. and Stockbridge, F. P. *Measure Your Mind: The Mentimeter and How to Use It.* New York: Doubleday, Page and Company, 1922. 349 p.

Washburne, C. W. "A Classified Scale for Measuring Intelligence," *Journal of Educational Psychology*, 10:309–22, September, 1919.

Yerkes, R. M., Bridges, J. W., and Hardwick, R. S. *A Point Scale for Measuring Mental Ability.* Baltimore: Warwick and York, 1915. 218 p.

Yerkes, R. M. and Foster, J. C. *A Point Scale for Measuring Mental Ability, 1923 Revision.* Baltimore: Warwick and York, 1923. 219 p.

II. Group Intelligence Tests

Chapman, J. C. "A Group Intelligence Examination Without Prepared Blanks," *Journal of Educational Research*, 2:777–86, December, 1920; 11:269–79, April, 1925.

Dearborn, W. F. and Lincoln, E. A. "Revising the Dearborn Intelligence Examinations," *Journal of Educational Psychology*, 14:39–46, January, 1923.

Franzen, Raymond. "Attempts at Test Validation," *Journal of Educational Research*, 6:145–58, September, 1922.

Gates, A. I. "The Correlations of Achievement in School Subjects with Intelligence Tests and Other Variables. Part III. Results from Grades IV to VIII," *Journal of Educational Psychology*, 13:223–35, April, 1922.

Goodenough, F. L. *Measurement of Intelligence by Drawings.* Yonkers, New York: World Book Company, 1926. 177 p.

Haggerty, M. E. "Intelligence Examination Delta 2," *Journal of Educational Psychology*, 14:257–77, May, 1923.

Hunt, Thelma. "The Measurement of Social Intelligence," *Journal of Applied Psychology*, 12:317–34, June, 1928.

Jordan, A. M. "The Validation of Intelligence Tests," *Journal of Educational Psychology*, 14:348–66, 414–28; September, October, 1923.

Kuhlmann, F. "The Kuhlmann-Anderson Intelligence Tests Compared with Seven Others," *Journal of Applied Psychology*, 12:545–94, December, 1928.

———. "Median Mental Age Method of Weighting and Scaling Mental Tests," *Journal of Applied Psychology*, 11:181–98, June, 1927.

MacPhail, A. H. *The Intelligence of College Students.* Baltimore: Warwick and York, 1924. 176 p.

Strasheim, J. J. *A New Method of Mental Testing.* Baltimore: Warwick and York, 1926. 158 p.

Thurstone, L. L. "A Cycle-Omnibus Intelligence Test for College Students," *Journal of Educational Research*, 4:265–78, November, 1921.

——— "Mental Tests for College Entrance," *Journal of Educational Psychology*, 10: 129–42, March, 1919.

CHAPTER XVI

PUPIL RATING

Introduction.—This chapter deals with various phases of the rating of pupils, their abilities and traits, not included under any of the school subjects or general intelligence. It is divided into three parts of which the first is devoted to the measurement of character and personality, using the terms in a very broad sense, the second to the measurement of what may be called school habits or school citizenship, and the third to the measurement of study. Of these three the first is the most important from the standpoint of the amount of attention it has received, but the other two probably have a closer connection with the achievements of pupils in school. There is ordinarily little demand that most teachers make formal ratings of pupils' characters and personalities. Nevertheless, this phase of measurement seems closely enough allied with the work of teachers that it should not be omitted from this volume. The measurement of the school-habits of pupils and of their methods of study should, on the other hand, be of immediate practical interest to every teacher, and should be closely connected with instructional and other activities in the regular subjects.

I. Character and Personality

This heading is used here to include tests, rating scales, and other measuring devices that have to do with a wide variety of traits and characteristics. Some of these belong very definitely to moral or ethical character, others have little if any connection with this, but are rather concerned with such phases of personality as aggressiveness or timidity, for example, sensitiveness to criticism or lack of sensitiveness, and so forth. Many of these characteristics are of psychological rather than educational interest, but on the other hand many of them are closely con-

nected with the usual work of the school. Furthermore, many of the measures of this sort have a more or less intimate connection with the problem of prognosis, especially vocational guidance. The more general of this type have been included here; those more narrowly intended for that purpose will be treated in Chapter XXII.

In this as in some other fields, the measuring instruments employed are of two chief varieties. One of these, which in this case is probably of lesser importance than the other, is the test; the other is the rating scale. Because of the fact that individuals commonly tend to respond to a test involving any moral element or any other that they think in any way reflects on their character or personality by giving what they believe are the correct responses rather than their responses in typical everyday situations, such tests are not valid measures of conduct. They measure what those tested know to do rather than what they actually do do. Some attempts, of which two or three will be mentioned in more detail later, have been made to avoid this by the use of situations in which the pupils do not know that they are being tested on such qualities, but the administration of these is usually rather difficult and time consuming, and there are also certain other arguments against their use, so that such means of measurement have not been developed to a satisfactory point. Therefore most of the efforts put forth in this field have had to do with developing rating scales and endeavoring to make them as objective, reliable, and valid as possible.

There are several possible different ways of grouping such scales. One is to divide them into ordinary rating scales and ranking scales, that is, scales that provide for ranking individuals in order rather than assigning them scores. There has been some argument as to which of the two methods is preferable, and apparently expert opinion is somewhere nearly evenly divided on the question. Each possesses certain advantages over the other, and likewise is subject to certain limitations. Perhaps ranking is better when the number of individuals to be dealt with is comparatively small, and rating when it is large. The reason for this is that it is decidedly difficult to arrange a large number of persons in order of rank.

Several more or less general means of improving judgment

or rating scores have been employed. One of these is to make up the scale for a particular trait by arranging in order from highest to lowest or best to worst a series of adjectives, descriptive phrases, or even longer statements. A common number is either five or seven. Sometimes these are arranged graphically, usually on a straight line, and raters are instructed to place a mark at the proper place on the line to indicate the degree of the quality which they believe the individual possesses. Sometimes scale steps are defined by statements of expected responses to situations on the theory that these are more concrete and meaningful than relatively abstract adjectives or descriptions. In some cases each trait to be rated is carefully defined so that there may be as little doubt as possible as to just what it is intended to include.

Despite these and other precautions there are several causes of lack of validity and reliability which it seems impossible to eliminate entirely. Perhaps the chief of these is the so-called "halo effect." This refers to the fact that a person rating another on a number of traits is influenced by his general impression or opinion of the individual to give similar ratings on all and not to differentiate carefully between the different characteristics in question. Possibly because of this, or possibly for some other reason, close acquaintance with the individual to be rated seems to lower the accuracy of the ratings given rather than to increase them. Another point of considerable difficulty is that most of the qualities of character or characteristics of personality which it is desired to rate are so complex, being made up of so many different elements, that they can scarcely be combined into a unified single rating. For example, suppose the trait of honesty is in question. Many, probably most, people are entirely honest in certain situations, but not so in others. Likewise if courage is being rated, most people are courageous in the face of certain dangers, but timid when they encounter others. Even with as exact definitions and delimitations as possible, this condition still holds. The comment should be made, however, that it is probably no more prevalent in this field than in the field of general intelligence, or any other mental trait. Although we commonly speak of people as having good or poor memories, for example, it is almost always true that a given individual has a relatively good memory for certain things that interest him, or

in which he is experienced, and a relatively poor memory for other things with which he has had little contact.

Many studies of the reliability of ratings given by different persons, or sometimes by the same persons at different times, have been made. The central tendency of the coefficients of reliability is about .55, although there are great variations from this figure. Shen,[1] for example, found average reliability coefficients of about .70 for ratings of scholarship, leadership, and intellectual quickness, and of .40 or below for resistance, adaptability, and impulsiveness. Another investigator[2] found considerably greater reliability in the rating of efficiency, originality, perseverance, and so forth, than in the cases of kindliness, cheerfulness, coöperativeness, and other similar traits. Rugg[3] concludes that character ratings can be made accurately enough for practical purposes if each is the average of three independent ratings made on our most objective types of scales, if the scales employed are comparable and equivalent, and if the persons doing the rating are well enough acquainted with the persons being rated. He goes on to state, however, that these conditions are practically unattainable in the public schools and that, therefore, the methods generally employed are not accurate enough to be of much value.

Before closing this discussion the matter of self-rating should be mentioned. This has not received as much attention as has rating by others. However, there are available fairly conclusive data to indicate that individuals rather uniformly tend to rate themselves too high in the traits that they consider of considerable importance, lower in others. The reliability of self-ratings is at least as high and probably somewhat higher than that of ratings given by others. Except for purposes of stimulating introspection, self-analysis, and improvement, they appear to be of little value.

The number of commercially available tests and scales that

[1] Shen, E. "The Reliability Coefficient of Personal Ratings," *Journal of Educational Psychology*, 16:232–36, April, 1925.

[2] Hollingworth, H. L. *Judging Human Character*. New York: D. Appleton and Company, 1922. p. 79.

[3] Rugg, H. O. "Is the Rating of Human Character Practicable?" *Journal of Educational Psychology*, 12:425–38, 485–501, November, December, 1921; 13:30–42, 81–93, January, February, 1922.

really merit use for the measurement of character and personality is decidedly limited. Partly for this reason, the following discussion will depart somewhat from the form generally used in the previous chapters on tests, and in addition to listing and describing several such instruments, will discuss in somewhat less formal style two or three of the outstanding attempts at measurement along this line that have not resulted in tests or scales available for general use.

Will-Temperament Tests
June E. Downey (1920)
Individual, Group

These are probably the best-known and most widely-used tests of what may be called the "will" or "emotional reactions." The individual test, which appeared in its preliminary form a year or two before the date given above, consists of exercises intended to measure the following twelve traits: speed of movement, freedom from load, flexibility, speed of decision, motor impulsion, reaction to contradiction, resistance, finality of judgment, motor inhibition, interest in detail, coördination of impulses, volitional perseveration. Most of the subtests call for writing under various conditions and directions. For example, the person being tested writes his name in his usual style and speed, as rapidly as possible and as slowly as possible; he imitates a model as rapidly as he can, and as exactly as he can; he attempts to disguise his writing as much as possible; he writes with his eyes closed while an obstruction is offered by the examiner, and so forth. Among the other exercises he checks the ones of pairs of contrasted words which he thinks more nearly describe him. These include such pairs as careful and careless, vain and modest, cheerful and gloomy, and so on. The group test is very similar, containing a number of identical exercises and others somewhat modified. The time required for either test is about one ordinary class period.

Miss Downey claims a reasonably high validity for these tests on a number of bases, most of which do not appear to be justified. Her reported correlation of .60 with judges' ratings of the same traits is scarcely high enough to be very significant, especially in

view of the fact that other investigators have reported lower ones. Correlations ranging from practically zero up to .63 with other similar tests are also not large enough to indicate high validity. A correlation with school marks of .32 and one with conduct of practically zero have also been found. A rather careful study of reliability based on several hundred cases has been made, each exercise being scored in a number of possible different ways. The first row below gives the range of the reliability measures for the different exercises and bases of scoring, and the second the average of all.

$$r = .05\text{–}.90,\ \text{P.E.}_{\text{meas.}},^{4}\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .04\text{–}.53,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .21\text{–}.66.$$

$$r = .65,\ \text{P.E.}_{\text{meas.}},^{4}\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .20,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .40.$$

Correlations between the individual and the group test ranging from slightly negative up to .90 have been reported. The average is only slightly above .30.

Despite the wide publicity which these tests have received and the numerous studies made of them, it is doubtful if they are of much practical value to the classroom teacher. On the other hand, they represent what is probably the most outstanding attempt to measure a phase of personality that is of considerable importance to teachers, personnel workers, and others. Attempts have been made to improve these tests both by actual revision and by new methods of scoring, but none of them have been shown to be better than the original.

World Book Company. Individual specimen set 25¢; $1.00 per 25; manual 20¢; record cards 35¢ per 25. Group specimen set 20¢; $1.40 per 25; manual 15¢ record cards 35¢ per 25.

References: Downey, J. E. *The Will-Temperament and Its Testing*. Yonkers, New York: World Book Company, 1923. 339 p.

———— "Testing the Will-Temperament Tests," *School and Society*, 16:161–68, August 5, 1922.

[4] Since so many different scoring units are used a general or average statement of the probable error of measurement would be meaningless, therefore none is given here.

May, M. A. "The Present Status of the Will-Temperament Tests," *Journal of Applied Psychology*, 9:29–52, March, 1925.

Uhrbrock, R. S. "An Analysis of the Downey Will-Temperament Tests," *Teachers College, Columbia University, Contributions to Education*, No. 296. New York: Bureau of Publications, Teachers College, Columbia University, 1928. 78 p.

Downey, J. E. and Uhrbrock, R. S. "Reliability of the Group Will-Temperament Tests," *Journal of Educational Psychology*, 18:26–39, January, 1927.

Ethical Discrimination Test

S. C. Kohs (1922)

This test consists of six parts that deal, respectively, with social relations, moral judgment, proverbs, definitions of moral terms, offense evaluation, and moral problems. The first requires the selection of the proper one of three possible responses to social situations or explanations of statements. In the second the worst term in each of twenty-five groups of five each must be indicated, most of the terms referring to more or less immoral acts or traits. The proverbs test presents twenty fairly common proverbs for each of which the best one of three explanations is to be indicated. In the fourth the proper one of four meanings for each of forty-five words is to be chosen. In the fifth a list of fifty words, most of which have some moral significance, is to be classified into six groups. The moral problems test calls for selecting the correct one of three given reasons for not doing each of ten stated acts. The total time required is about forty minutes.

This is among the more elaborate of the tests that attempt to measure moral or ethical qualities, but the results are in a very high degree subject to the criticism that they measure knowledge and not practice.

C. H. Stoelting Company. $6.75 per 25, $2.00 per 100; manual 65¢.

Reference: Kohs, S. C. "An Ethical Discrimination Test," *Journal of Delinquency*, 7:1–15, January, 1922.

X-O Tests
S. L. Pressey
Forms A, B

These tests are also frequently referred to as "tests of the emotions." Test A, intended chiefly for adults, consists of four subtests. In the first are twenty-five lists of five words each, from which every word whose meaning is unpleasant is to be crossed out and also the one most unpleasant in each list indicated. In Subtest II are twenty-five words each followed by a list of five with directions to indicate all of the five connected with the first word and also the one most closely connected with it. The third also has twenty-five groups of five words each, in which all that refer to something wrong are to be indicated and the one most wrong in each group marked. The last subtest is similar except that things the person being tested has worried or felt nervous about are to be marked. Form B is not a duplicate of A, but instead is a simpler and somewhat expurgated form with the fourth subtest omitted, and is intended for use in elementary and high school. Most of the words included in both tests are such as have definite moral implications or else are likely to arouse fairly strong emotions. For example, one list of five in the first subtest is "cruel shirt favorite laughter crawl," another, "distance slippery cannibal assault persecute." Pupils are to be allowed as much time as needed, but the time is recorded. Presumably thirty minutes or thereabouts are sufficient.

The test is scored in terms of the difference of an individual's responses from the normal or average. A number of so-called "jokers," that is, words that should arouse no emotional responses, are inserted. If too many of these are crossed out it is

Subtest	*Test B-High School*				*Test A College*
	IX	*X*	*XI*	*XII*	
I	70	67	72	58	41
II	24	31	38	38	55
III	40	46	61	68	73
IV	—	—	—	—	46
Total	137	165	175	170	230

likely that either the directions were not understood or there was not satisfactory coöperation. Median numbers of words crossed out, apparently based on a relatively small number of cases, are given at the bottom of page 421.

The average numbers of words marked that are not the same as the ones most usually marked, are as follows:

Subtest	*Test B-High School*				*Test A College*
	IX	*X*	*XI*	*XII*	
I	17	16	15	13	11
II	18	18	16	15	10
III	18	17	16	16	13
IV	—	—	—	—	15
Total	53	51	48	46	47

Although these facts do not appear to have much immediate practical value in connection with school achievement, it has been found that a correlation of .43 exists between scores on this test and school marks. In the same experiment a correlation of practically the same amount was found between intelligence test scores and school marks, so that apparently this is as good as a group intelligence scale for predicting success in school.

C. H. Stoelting Company. $1.00 per 25, $3.00 per 100.

References: Pressey, S. L. "A Group Scale for Investigating the Emotions," *Journal of Abnormal Psychology and Social Psychology*, 16:55–64, April, 1921.

Thompson, Jr., L. A. and Remmers, H. H. "Some Observations Concerning the Reliability of the Pressey X-O Test," *Journal of Applied Psychology*, 12:477–94, October, 1928.

Tests of Trustworthiness

P. F. Voelker (1921)

Series I, II; Tests 1–10 of each

These tests are among those referred to at the beginning of this section as not being available for general use in the sense of being for sale by commercial publishers,[5] but as meriting inclu-

[5] Most of the tests are described with sufficient completeness that they can be reproduced if desired.

sion because of the methods employed and results attained. Furthermore, these series deserve mention because they represent one of the first comprehensive attempts at measurement along this line.

Series I and II were not duplicates, although intended to measure the same general trait. The former included ten tests as follows:

1. *Overstatement.* Opportunity was offered to agree with an overstatement of a school mark.

2. *Suggestibility.* Those tested were contradicted to see if they would agree with the contradiction.

3. *Willingness to accept help.* Help was offered in the solution of puzzles that were to be solved independently.

4. *Borrowing errand.* The subject was sent to borrow something and promise to return it by a given time. Later he had the opportunity to keep his promise or fail to do so.

5. *Purchasing errand.* By pre-arrangement the subject received overchange and his disposition of it was determined.

6. *Tip.* A tip was offered for a trifling courtesy.

7. *Push button.* The subject was instructed to push a button at regular intervals, but a number of distracting objects were left in his presence.

8. *Distraction.* All a's in uninteresting material were to be copied, and following that all a's in an illustrated book.

9. *Profile board.* After having solved a profile board puzzle correctly with the eyes open, the subject was required to do three similar ones with eyes closed and report his success or failure. (There was only about one chance in four thousand that all three could be done correctly, so that a report of three successes was almost certainly false.)

10. *Tracing and opposites.* After giving the opposites of a list of words, subjects had the opportunity of cheating by correcting their original answers under such conditions that the corrections would be discovered.

Series II contained ten more or less similar tests.

As is evident from the brief descriptions of the tests given, about half of them, in addition to portions of others, had to be given individually. The time consumed was, therefore, considerable. The results were correlated with a number of other meas-

ures and otherwise manipulated. Correlations of from .30 up to .93 were found with the judgments of teachers and scout leaders. It was also shown that scores tended to increase according to the amount of scout training received, although the correlation was far from perfect. Comparison of the results of the two series indicated fairly high reliability, although it was lower than that on most widely used achievement tests. Judging from increase in scores between the first and second administration of the tests, general intelligence played little part in the improvement of trustworthiness.

Reference: Voelker, P. F. "The Function of Ideals and Attitudes in Social Education; An Experimental Study," *Teachers College, Columbia University, Contributions to Education*, No. 112. New York: Bureau of Publications, Teachers College, Columbia University, 1921, Chapter V.

Character Education Inquiry Tests
Hugh Hartshorne, M. A. May, et al. (1926)

As part of the Character Education Inquiry being conducted by the Institute of Educational Research of Teachers College, Columbia University, and the Institute of Social and Religious Research, Hartshorne and May have had charge of the development of a series of tests. Few if any of them are available for general use, and because of this fact coupled with the relatively large amount of space that would be required to describe them at all completely, they will not be dealt with in detail. The tests included several intended to measure certain character traits without disclosing to the pupils that they were doing so, others intended to measure knowledge of right and wrong rather than habits of conduct, and others dealing with various related matters. Several of them were vocabulary tests containing words that seem to have some connection with morality or character and calling for the giving of opposites, synonyms, and so forth. As examples of the words included, the following may be given: "friend," "help," "debase," "loath," "cheating," and "loyalty." One true-false test was composed of statements dealing with cause and effect, usually of a more or less moral character, another with duties, and another with good manners. Several

multiple-answer tests called for the classification of statements of acts as "right," "excusable" or "wrong," or as "cheating," "lying," "stealing," "something else wrong," or "not wrong at all," and so forth. Several tests dealt with such common happenings as a boy accidentally breaking a street lamp, a boy stealing some things from a five and ten cent store, and so on, calling for responses as to what was the right thing to do, what was likely to occur, what important consequences would follow, and so on. One, designed to test perseverance or willingness to follow directions, first determined the pupil's rate in accuracy of solving simple examples in arithmetic, and then how well he would continue to do so when distractions in the form of interesting stories, pictures and puzzles were kept before him. Another set of tests was similar to one used by Voelker, requiring pupils to draw certain lines or make certain marks with eyes closed, these being so difficult that it was practically impossible to succeed without cheating by opening the eyes.

This work of Hartshorne and May is without doubt the most ambitious and important attempt to measure character and conduct so far carried out. Despite this fact, the results are useful rather as the basis for future investigations than as yielding conclusive information of any sort. Most of the tests are given to enough pupils that reasonably satisfactory norms, intercorrelations, reliability data, and so forth, were secured. On the whole, the tests compare favorably in reliability with those in the school subjects, a number of them yielding coefficients around .90. In some cases results were secured which afford fairly good evidence as to the relative importance of different groups of persons as sources of children's knowledge of right and wrong. These indicate that parents are most important, friends next, group leaders next but low, and both public and Sunday school teachers of practically no importance.

It is possible to secure copies of some of the tests from the Institute of Educational Research, or perhaps from the authors.

References: Hartshorne, Hugh, May, M. A. et al. "Testing the Knowledge of Right and Wrong," *Religious Education*, 21:63–76, 239–52, 413–21, 539–54, 621–32, February, April, August, October, December, 1926; 22: May, 1927. Also in: *Religious Education Association Monograph*, No. 1 Chicago: Religious Education Association, 1927. 72 p.

Hartshorne, Hugh and May, M. A. "Studies in Deceit," *Studies in the Nature of Character*, Vol. I. New York: The Macmillan Company, 1928. Book One, 414 p.; Book Two, 306 p.

Trait Index L

Elizabeth H. Morris (1929)

This is a self-administering trait index that may be given either individually or to groups. It consists of six sections, each of which provides for self rating of a slightly different kind. In the first forty-six items, such as "studying," "travel," "people between one and six years of age," "talking to strangers," "having high ideals," and so forth, are to be marked as belonging to one of five groups ranging from "like very much" to "dislike very much." Section II lists fourteen possible comments teachers might make to different kinds of pupils, and calls for each being marked according to which one of six types of pupils it fits best. In the third section are seventeen situations likely to occur in the schoolroom that are to be marked as "amusing," "embarrassing," "necessitating firm control," "interesting," or "necessitating correction of mistake." Section IV contains twelve more or less similar situations, but gives four possible responses to each from which the best one is to be chosen. In the next section are forty-one statements about school work, pupils, and so forth, to be labelled with one of five symbols that indicate degrees of truth, ranging from "always true" to "never true." In the last section seven situations are presented with six possible attitudes toward each. The individual is to indicate the one of the six most natural to him. There is no time limit, but it is stated that most people require from forty to sixty minutes.

This instrument represents an attempt to measure feeling or attitude rather than knowledge. The degree to which this succeeds evidently depends in this, as in nearly any other case, upon how far the examiner and general conditions of testing are able to induce pupils to be frank in their answers. On the whole, the wording of the directions and content of the exercises are such as to indicate that this is among the best available instruments of this sort. Results show some correlation between ratings on this and probable success in teaching, also a reasonably high reliability. No norms are given for high-school pupils, but for college freshmen,

probably a slightly superior group, the average score is about 105. The author states that a high score has greater favorable significance than a low score has unfavorable. Provision is made for a graphic profile in which scores on certain sections are taken as measures of particular phases of personality or character traits.

Public School Publishing Company. Sample set 15¢; $1.50 per 25.

II. School Habits and Attitudes

Although some of the measuring instruments, and especially the last one, referred to in Section I of this chapter, contain exercises dealing with school attitudes and characteristics, there are others so completely devoted to this purpose that they appear to merit separate consideration. Most of these rating scales have been developed by single schools or school systems for their own use, perhaps occasionally written up in periodical articles or elsewhere, but not given much publicity nor made available for others. Indeed, there are only two or three instruments of this type published commercially that appear to possess sufficient merit to warrant describing them in this section.

Although the same situation exists in practically any field of testing, it is probably more true in this than in most that there are benefits to be derived by a group of teachers from constructing their own rating scale. A scale of this type will not yield as objective scores as a good test, but if teachers make their own they will have clearer and more definite ideas as to just what the terms used in it mean and, therefore, will be able to rate more accurately by it.

The uses to which such scales are put are numerous. Probably the most common one is in connection with pupils whose work is unsatisfactory in the attempt to discover the reasons why this condition exists and what remedial measures may be taken. More and more schools, however, are recognizing that not only pupils who are failing or in danger of failing should receive help of this sort, but that all pupils are entitled to it. Such scales are also being used by higher institutions, chiefly in connection with the admission of students. Still another situation in which they are largely employed is in the recommendation of individuals for various

positions, especially as teachers. Since, however, these latter purposes pertain chiefly to the college or university and not to the high school and, furthermore, since general discussions of placing and classifying pupils, of educational and vocational guidance, and so forth, will occur in later chapters, no further discussion will be given here.

Rating Scale for School Habits[6]

E. L. Cornell, W. W. Coxe, and J. S. Orleans (1927)

This is a simple little scale dealing with nine qualities as follows: attention, neatness, honesty, interest, initiative, ambition, persistence, reliability, and stability. Under each three degrees, extremely low, average, and extremely high, are briefly described. A line reaches from the first past the second to the third, and teachers are to mark the point on each line which they think best represents the amount of the trait possessed by the pupil being rated. The back of the sheet provides for recording brief notes as to how teachers gained their knowledge of the children rated. After the ratings have been made it is suggested that a line be drawn connecting them, thus giving a profile or graph that brings out the pupil's strong and weak points.

Correlations between .55 and .75 have been found for ratings on this scale and first term high-school marks. The authors also make the general statement that measures of school habits correspond about twice as closely with school marks as do intelligence test scores. This scarcely appears justified, however, on the basis of the correlations just quoted.

World Book Company. 50¢ per 25.

Joliet Rating Scale

This scale is one of the best of those prepared by schools for their own use with which the writer is familiar, therefore it will be described here as representative of such scales. It deals with six qualities: sustained application, ability to organize, promptness, accuracy, leadership ability, and social qualities. Five degrees of each are described by three phrases or statements. For example,

[6] This is sometimes called the New York Rating Scale for School Habits.

the highest degree of sustained application is described as follows:

"Always busy; Work purposeful. Sees job through to completion. Does more than is required."

The lowest degree of this is:

"Seldom or never busy; Absence of purpose. Never finishes job well. Gets out of as much as possible; 'Gets by' with as little as possible."

The scale is intended to be used by teachers and the results employed by the personnel office of the school. Teachers are told that for a sufficiently large and normal group of pupils the distribution of ratings should approximate a normal curve, with per cents of 7, 23, 40, 23, and 7 in the five classes.

Joliet Township High School.

III. Study

A number of the rating scales of the type described in the last section provide for rating pupils either directly upon their study habits or upon traits or qualities closely connected therewith. In addition to these, however, a few attempts have been made to construct standardized tests of study habits or ability. Only one test of this type suitable for use in high school appears to be actually for sale at present. This will be described and in addition one series of such tests never commercially available, but indicative of a type of testing for which there is need.

The paucity of tests of pupils' study habits is, however, not quite as great as is suggested by the preceding paragraph. Most of the studying done by high-school pupils is chiefly reading, so that most, if not all, reading tests, although not so labelled, are in a sense study tests also. This is especially true of such tests as the Van Wagenen Reading Scales in English Literature, History, and General Science, the Iowa Comprehension Tests, and others described in the section of Chapter V on reading tests which deal particularly with the kind of material high-school pupils must read in their several subjects. This will be made more evident by the fact that the series of diagnostic study tests described in this section might almost, if not quite, as well be called reading tests. Fairly satisfactory diagnoses of pupils' study habits can, therefore, be made by combining results of available tests with observa-

tion of pupils while at work, and perhaps some questioning of them as to their mental activities during that time.

Study Outline Test
F. D. McClusky and E. W. Dolch (1926)
Test 1, 2, 3

Each test consists of a selection, the same one being used for all three, to be outlined according to a simple system explained in the directions. In Test 1 no help is given; in Test 2 certain words that should be of assistance in outlining are inserted; and in Test 3 both words and numbers are contained in the text. It is intended that Test 1 be given first. Pupils who pass it evidently do not need the other two. Those who fail upon it should later be given Test 2, and those who fail thereon, after a period of time should take Test 3. After further instruction, 2 and 1 may then be taken again. From ten to fifteen minutes is sufficient time.

Norms for these tests are given as follows:

Test	*High School*				*University*			
	I	*II*	*III*	*IV*	*I*	*II*	*III*	*IV*
1	3.8	3.8	7.1	7.9	11.4	11.4	12.6	13.8
2	5.4	5.7	8.1	8.7	12.1	12.8	13.1	13.5
3	6.0	6.1	10.2	11.2	15.2	15.2	15.4	16.5

Public School Publishing Company. Sample set 10¢; 75¢ per 25 of Test 1, 15 of Test 2, and 10 of Test 3.

Reference: McClusky, F. D. and Dolch, E. W. "A Study Outline Test," *School Review*, 32:757–72, December, 1924.

Diagnostic Study Tests
Bureau of Educational Research, University of Illinois (1922)
Tests I, II, III, IV, V

Although these tests were never standardized and are not now commercially available,[1] it seems worth while to describe them

[1] A small supply of each is still on hand at the Bureau of Educational Research and limited quantities will be sent to those requesting them.

briefly because they appear to measure certain significant abilities and to suggest how teachers can construct similar tests for their own use. Each test requires the reading of a number of selections of typical high-school content, but the purpose varies from one test to another. In the first pupils read to secure answers to certain questions that they already know. In the second test they read paragraphs with a view to remembering them well enough to answer questions that they will encounter later when they cannot look back at the selections. In Test III are paragraphs leading to conclusions followed by a number of statements of which the relevant ones are to be checked and problems followed by a number of statements of fact of which all needed for the solution are to be indicated. In Test IV each selection is followed by a number of statements. The one that best expresses the central idea of each paragraph is to be marked. The fifth deals with outlining. Each selection is followed by a number of statements to be properly organized to form an outline. Although these tests do not cover all types of study, yet they represent a beginning in this direction and were found to yield information of more or less value in dealing with pupils, especially, of course, those not doing work of high merit.

Reference: Monroe, W. S. and Mohlman, D. K. "Errors Made by High-School Students in One Type of Textbook Study," *School Review*, 31:36–47, January, 1923.

BIBLIOGRAPHY

I. Character and Personality

Athearn, W. S., et al. "Measurements and Standards in Religious Education," *Indiana Survey of Religious Education*, Vol. II. New York: George H. Doran Company, 1923. 532 p.

Cady, V. M. "The Estimation of Juvenile Incorrigibility; A Report of Experiments in the Measurement of Juvenile Incorrigibility by Means of Certain Non-Intellectual Tests," *Journal of Delinquency Monograph*, No. 2. Whittier, California: Whittier State School, 1923. 140 p.

———. "The Psychology and Pathology of Personality, A Summary of Test Problems and Bibliography of General Literature," *Journal of Delinquency*, 7:225–48, September, 1922.

Clark, W. W. "The Measurement of Social Attitudes," *Journal of Applied Sociology*, 8:345–54, July–August, 1924.

———. "Whittier Scale for Grading Juvenile Offenses," *California*

Bureau of Juvenile Research Bulletin, No. 11. Whittier, California: Whittier State School, 1922. 8 p.

Filter, R. O. "An Experimental Study of Character Traits," *Journal of Applied Psychology*, 5:297–317, December, 1921.

Furfey, P. H. "Tests for the Measurement of Non-Intellectual Traits," *Catholic University Educational Research Bulletin*, Vol. 3, No. 8. Washington, D. C.: Catholic Education Press, 1928. 35 p.

————. "Tests for Personality Traits—A Review of the Literature," *Catholic Education Review*, 25:614–21, December, 1927.

Gorham, Donald R. and Brotemarkle, R. A. "Challenging Three Standardized Emotional Tests for Validity and Employability," *Journal of Applied Psychology*, 13:554–88, December, 1929.

Hart, H. N. "Progress Report on a Test of Social Attitudes and Interests," *University of Iowa Studies in Child Welfare*, Vol. 2, No. 4. Iowa City: University of Iowa, 1923. 40 p.

Howell, G. D. "A Description of Some Tests for Traits Other than Intelligence," *Chicago Schools Journal*, 6:47–50, October, 1923.

Hughes, W. H. "General Principles of Rating Trait Characteristics," *Educational Research Bulletin, Pasadena City Schools*, 3:3–11, February–March, 1925.

————. "A Rating Scale for Individual Capacities, Attitudes, and Interests," *Journal of Educational Method*, 3:56–65, October, 1923.

————. "Some Strong Points and Some Weaker Points in Honor Students," *Educational Research Bulletin, Pasadena City Schools*, 2:3–9, November, 1923.

Jones, Vernon. "Ideas on Right and Wrong Among Teachers and Children," *Teachers College Record*, 30:529–41, March, 1929.

McGrath, M. C. "A Study of the Moral Development of Children," *Psychological Monograph*, Vol. 32, No. 2. Princeton, New Jersey: Psychological Review Company, 1923. 190 p.

Marston, L. R. "The Emotions of Young Children," *University of Iowa Studies in Child Welfare*, Vol. 3, No. 3. Iowa City: University of Iowa, 1925. 99 p.

Mary, Sister, Gannon, M. A., and Moloney, H. M. "An Extension of the Moral Information Tests," *Catholic University Educational Research Bulletin*, Vol. 3, No. 5. Washington, D. C.: Catholic Education Press, 1928. 31 p.

May, M. A. and Hartshorne, Hugh. "First Steps Toward a Scale for Measuring Attitudes," *Journal of Educational Psychology*, 17:145–62, March, 1926.

————. "Objective Methods of Measuring Character," *Pedogogical Seminary*, 32:45–67, March, 1925.

————. "Personality and Character Tests," *Psychological Bulletin*, 23:395–411, July, 1926.

May, M. A., Hartshorne, Hugh, and Welty, Ruth. "Personality and Character Tests," *Psychological Bulletin*, 24:418–35, July, 1927.

————. "Personality and Character Tests," *Psychological Bulletin*, 25:422–43, July, 1928.

Ream, M. J. "Group Will-Temperament Tests," *Journal of Educational Psychology*, 13:7–16, January, 1922.

Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction*. Yonkers, New York: World Book Company, 1927, p. 312–17.

Sunné, Dagny. "Personality Tests: White and Negro Adolescents," *Journal of Applied Psychology*, 9:256–80, September, 1925.

Symonds, P. M. "The Measurement of Conduct," *Measurement in Secondary Education*. New York: The Macmillan Company, 1927, Chapter XVI.

———. "The Present Status of Character Measurement," *Journal of Educational Psychology*, 15:484–98, November, 1924.

———. "A Social Attitudes Questionnaire," *Journal of Educational Psychology*, 16:316–22, May, 1925.

Terman, L. M. "The Physical and Mental Traits of Gifted Children," *Twenty-third Yearbook of the National Society for the Study of Education*, Part I. Bloomington, Illinois: Public School Publishing Company, 1924, p. 155–67.

Uhrbrock, R. S. and Downey, J. E. "A Non-Verbal Will-Temperament Test," *Journal of Applied Psychology*, 11:95–105, April, 1927.

Woodrow, Herbert. "A Picture-Preference Character Test," *Journal of Educational Psychology*, 17:519–31, November, 1926.

"Written Tests of Honesty (Integrity)," *Public Personnel Studies*, 7:98–106, July, 1929.

II. School Habits and Attitudes

Chassell, C. F., Upton, S. M., and Chassell, L. M. "Short Scales for Measuring Habits of Good Citizenship," *Teachers College Record*, 23:52–79, January, 1922.

Hill, E. L. "A Citizenship Rating Scale," *Education*, 47:362–71, February, 1927.

Upton, S. M. and Chassell, C. F. "A Scale for Measuring the Importance of Habits of Good Citizenship," *Teachers College Record*, 20:36–65, January, 1919. *Also in: Teachers College Bulletin*, Series 12, No. 9. New York: Teachers College, Columbia University, 1921. 44 p.

Van Buskirk, Luther. "Measuring the Results of Physical Education," *Journal of Educational Method*, 7:221–29, February, 1928.

———. "A Rating Scale for Use in Citizenship Training," *High School Teacher*, 3:18–19, January, 1927.

III. Study

Daringer, H. F. "An Objective Measure of Ability to Make Topical Outlines," *Journal of Educational Psychology*, 20:112–18, February, 1929.

Symonds, P. M. "A Studiousness Questionnaire," *Journal of Educational Psychology*, 19:152–67, March, 1928.

CHAPTER XVII

TEACHER RATING

Introduction.—Although the ordinary high-school teacher does not rate teachers, except perhaps by giving self-ratings, yet she is enough concerned in the rating of teachers that it seems worth while to devote a short chapter to its consideration. In practically all schools teachers are rated by superintendents, principals, supervisors, or other officials. In many, perhaps most, cases the rating is informal, no definite scheme of rating being employed, but instead a general impression formed by the supervisory official. In many systems, however, more or less formal scales or score cards are in use. Frequently they have been worked out by local officials and are peculiar to the single school or system in which they are employed. Indeed, most of the scales are probably either of this sort or else are used by a very small number of schools in contrast to the comparatively few that have received wide use. The number of such instruments offered for sale or otherwise available to anyone desiring them is not great, but a much larger number may be found described in periodical articles and elsewhere. Several of this type will be included among those described in this chapter, since its main purpose is not to present certain instruments in the hope that their use will be encouraged, but rather to indicate what some of the best scales are like and the different types thereof. This, it is hoped, will encourage those who use such scales to develop their own, patterning them more or less after such as are described here, and also encourage teachers to study such scales both in order that they may improve their efficiency by directing conscious attention to the various factors supposed to constitute good teaching, and that they may improve the scales actually employed by criticism of them and participation in their construction.

Of the two purposes just stated the first is undoubtedly the most important. It should include not merely a more or less impersonal

consideration of what teachers should do to merit high ratings, but direct personal self-ratings from time to time. A few of the scales are definitely intended for this purpose, but practically all may be employed in this manner with gain to the person doing so. In teaching, as in any other occupation, self-analysis and self-criticism is one of the most helpful means to improvement.

Reference was made above to the fact that there are various types of teacher rating scales and score cards. From one standpoint they may be classified as those for practice or student teachers, for classroom teachers actually in service, and a very few for the rating of principals and supervisors. From the standpoint of form or points covered, the varieties are numerous. Some are very short, containing mainly a few general points, whereas others go into great detail, a few even running into hundreds of items. Some are mere lists of questions or points, with no values or weights attached, whereas others have been carefully weighted. Some are general, covering in so far as possible the whole scope of teachers' activities and frequently personality and so forth as well, whereas others are limited to actual instruction, or even only certain phases thereof. Among the few more or less standardized tests in education there are one or two which do not deal with the content of particular courses as usually given that may well be considered as instruments for general teacher rating.

Scale for Rating Teachers
T. H. Schutte (1923)

This scale is a revision of an earlier scale by the same author, known as the Moorhead Hundred Point Scale for Rating Teachers. In addition to blanks for training, experience, and certain general information, this scale consists of five divisions that deal with personal and social qualities, coöperative qualities, leadership, scientific and professional attitude, and teaching ability. These contain a total of eighty-six questions, some of which are subdivided, dealing with one hundred points. The questions deal with such matters as intellectual capacity, health, self-reliance, enthusiasm, attitude toward pupils, attention to heat, light, and ventilation, use of English, making assignments, teaching of study habits, motivation of work, and so on. Each is followed by five degrees of

CHAPTER XVII

TEACHER RATING

Introduction.—Although the ordinary high-school teacher does not rate teachers, except perhaps by giving self-ratings, yet she is enough concerned in the rating of teachers that it seems worth while to devote a short chapter to its consideration. In practically all schools teachers are rated by superintendents, principals, supervisors, or other officials. In many, perhaps most, cases the rating is informal, no definite scheme of rating being employed, but instead a general impression formed by the supervisory official. In many systems, however, more or less formal scales or score cards are in use. Frequently they have been worked out by local officials and are peculiar to the single school or system in which they are employed. Indeed, most of the scales are probably either of this sort or else are used by a very small number of schools in contrast to the comparatively few that have received wide use. The number of such instruments offered for sale or otherwise available to anyone desiring them is not great, but a much larger number may be found described in periodical articles and elsewhere. Several of this type will be included among those described in this chapter, since its main purpose is not to present certain instruments in the hope that their use will be encouraged, but rather to indicate what some of the best scales are like and the different types thereof. This, it is hoped, will encourage those who use such scales to develop their own, patterning them more or less after such as are described here, and also encourage teachers to study such scales both in order that they may improve their efficiency by directing conscious attention to the various factors supposed to constitute good teaching, and that they may improve the scales actually employed by criticism of them and participation in their construction.

Of the two purposes just stated the first is undoubtedly the most important. It should include not merely a more or less impersonal

consideration of what teachers should do to merit high ratings, but direct personal self-ratings from time to time. A few of the scales are definitely intended for this purpose, but practically all may be employed in this manner with gain to the person doing so. In teaching, as in any other occupation, self-analysis and self-criticism is one of the most helpful means to improvement.

Reference was made above to the fact that there are various types of teacher rating scales and score cards. From one standpoint they may be classified as those for practice or student teachers, for classroom teachers actually in service, and a very few for the rating of principals and supervisors. From the standpoint of form or points covered, the varieties are numerous. Some are very short, containing mainly a few general points, whereas others go into great detail, a few even running into hundreds of items. Some are mere lists of questions or points, with no values or weights attached, whereas others have been carefully weighted. Some are general, covering in so far as possible the whole scope of teachers' activities and frequently personality and so forth as well, whereas others are limited to actual instruction, or even only certain phases thereof. Among the few more or less standardized tests in education there are one or two which do not deal with the content of particular courses as usually given that may well be considered as instruments for general teacher rating.

Scale for Rating Teachers
T. H. Schutte (1923)

This scale is a revision of an earlier scale by the same author, known as the Moorhead Hundred Point Scale for Rating Teachers. In addition to blanks for training, experience, and certain general information, this scale consists of five divisions that deal with personal and social qualities, coöperative qualities, leadership, scientific and professional attitude, and teaching ability. These contain a total of eighty-six questions, some of which are subdivided, dealing with one hundred points. The questions deal with such matters as intellectual capacity, health, self-reliance, enthusiasm, attitude toward pupils, attention to heat, light, and ventilation, use of English, making assignments, teaching of study habits, motivation of work, and so on. Each is followed by five degrees of

excellence, ranging from "E" or far below average, up to "A" or far above average, and also a very small place for additional comments. Directions provide for approximate adherence to a normal distribution. As Schutte himself points out, this scale, in common with many others, has one marked defect. Teacher activities rather than pupil activities are emphasized. Despite this and certain minor points, the writer believes it is among the few best available scales.

World Book Company. $1.00 per 25.

South Dakota Score Cards for Rating Teachers

W. A. Cook (1921)

Cards B, C

Card B is intended for use in systems of from six to twelve teachers, and Card C in larger ones. The items on each are grouped under the following seven main headings: scholarship, professional spirit and training, teaching ability, school management, material conditions, personal equipment and appearance, extramural efficiency. Card B has a total of twenty-eight points listed under these heads, and C of thirty-six. Teachers are to be rated in one of five classes, ranging from excellent to failure on each point. It is suggested that this be done by having in mind one very good, one average, and one very poor teacher, and that the distribution of scores approximate normality.

Bureau of Educational Research, Northern Normal and Industrial School. Not available in quantity.

University of California at Los Angeles Rating Scale for Practice Teaching

C. W. Waddell (1928)

Although this scale is particularly intended for use with practice teachers, an examination of it does not reveal any important reasons why it should not be employed with experienced teachers also. There are four main divisions dealing with personal qualities, academic and professional background, classroom management, and teaching skill. Each contains several subdivisions with

several points under each. The total number of points is about seventy. The scale provides for rating each individual at one of five degrees ranging from failure to superior on each point. The passing ratings are to follow approximately a suggested distribution that is not quite normal. The scores on the various points are combined to give one on each of the four chief divisions, and these in turn to give one general rating.

C. W. Waddell. 5¢ per copy; reduction on 100 or more.

Reference: Waddell, C. W. "A New Rating Scale for Practice Teaching," *Journal of Educational Method*, 8:214–19, January, 1929.

Checking List and Standards for Supervision of High School Instruction

F. W. Johnson (1924)

This instrument is similar to many teacher-rating scales except that it deals with the matter from the standpoint of the supervisor and not of the teacher. There are three main divisions dealing with classroom management, selection and arrangement of subject matter, and the recitation, with a total of more than fifty questions under them. For the benefit of the supervisor these are accompanied by standards, usually one or two statements for each question, stating what the ideal is. Provision is made for three ratings on each point, A, B, or C, of which A and C are each to be given to about 20 per cent of teachers, and B to the remaining 60 per cent. There is also a space for remarks after each question, and a space for noting most commendable and weakest features.

Bureau of Publications. 10¢ per copy.

Scales for the Rating of Teaching Skill

L. J. Brueckner (1927)

This series of scales differs considerably from most of the rating schemes used. Following Courtis, Brueckner classifies methods of teaching as being of four types: compulsion, teacher preparation, motivation, and purposing. Under each of these types he describes in several paragraphs each a series of nine recitations. These are arranged in order from best to worst and each given a scale value.

Abbreviated scales consisting of only five of each set of nine are also suggested. The scales were prepared by having over fifty descriptions of lessons of each of the four types rated by more than one hundred judges and then selecting nine from each set as nearly equally spaced from one another as possible. In order to have uniform subject matter, all the descriptions deal with elementary school geography, but despite this fact they may be used in rating high-school teachers also.

Although these scales were not prepared for general distribution, it is probable that copies of the publication listed below, which contains them in complete form, can be secured.

Reference: Brueckner, L. J. "Scales for the Rating of Teaching Skill," *University of Minnesota, Educational Research Bulletin*, Vol. 30, No. 12. Minneapolis: University of Minnesota, 1927. 28 p.

Classroom Procedure Tests (1928)

General Method—Douglas Waples and W. C. Reavis
English Composition—H. A. Anderson and Douglas Waples
English Literature—I. C. Poley and Douglas Waples
History and Other Social Studies—H. E. Wilson and Douglas Waples
Mathematics—E. R. Breslich and C. A. Stone
Natural Science—H. A. Cunningham and Douglas Waples

Form A of each

Each of this series of tests contains ten subtests presented in a page apiece. The points dealt with differ in the different tests. Those in the General Methods Test, of course, have to do with situations or problems likely to arise in any subject, whereas those in the others are largely peculiar to the separate subjects. For example, the subtests of the English Composition Test deal with such topics as introducing the course, proposing topics for compositions, indicating pupils' errors, discouraging the use of slang, and so forth. At the end of each test is a data blank calling for a number of items of information about training, experience, and so forth, and also a summary statement of the answers to the tests. These answers in each case consist of selecting the most and the

least efficient of a number of given procedures for use in the given situation.

Most of the situations or problems presented in the tests are very practical, being such as are likely to occur to teachers in service. The suggested procedures in most cases are those most commonly employed by teachers, or which teachers would perhaps most readily think of. In some cases at least there could well be more procedures suggested than the five given. The authors state that they have collected considerable data regarding norms, reliability, and validity, but that these will not be published until other material necessary to a satisfactory interpretation has been collected and studied.

University of Chicago Press. Complete set $1.00; 10¢ per copy, $1.25 per 25; manual 10¢.

Aptitude Tests for Elementary Teachers

J. E. Bathurst, F. B. Knight, G. M. Ruch, and Fred Telford

(1925)

Set 1[1]

This test has six subtests headed as follows: Professional Judgment, Theory and Practice of Teaching, Reading Comprehension, Social Information, School and Class Management, and Professional Information. These call for a total of two hundred fifty responses to multiple-answer and true-false exercises.

Although this is labelled as being for elementary teachers, the content is such that most of it is appropriate for high-school teachers as well. The large number of elements and the wide range they cover make the scores on this test rather good indications of teachers' knowledge of educational methods and related topics. Reliability data are as follows:

$$r = .80,\ \text{P.E.}_{\text{meas.}} = 6,\ \frac{\text{P.E.}_{\text{meas.}}}{M} = .05,\ \frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .30.$$

Norms for about five hundred fifty high-school teachers are:

[1] Set 1 was originally published as Test 1—Aptitude, of Professional Tests for Elementary Teachers. This series also includes a placement test dealing with knowledge of teaching methods in reading, arithmetic, spelling, and writing.

Percentile	*5*	*10*	*25*	*50*	*75*	*90*	*95*
Norm	73	81	98	115	131	148	157

Bureau of Public Personnel Administration. Specimen set 60¢; $3.00 per 25, $10.00 per 100, $45.00 per 500, in larger quantities 8¢ per copy.

BIBLIOGRAPHY

Abbott, Allan. "Tests for English Teachers," *English Journal*, 12:663–71, December, 1923.

Brandenburg, G. C. and Remmers, H. H. "Rating Scales for Instructors," *Educational Administration and Supervision*, 13:399–406, September, 1927.

Brown, E. J. "A Self-Rating Scale for Supervisors," *American School Board Journal*, 77:36–37, 127, August, 1928.

Carrigan, R. A. "The Rating of Teachers on the Basis of Supervisory Visitation," *Journal of Educational Method*, 2:48–55, October, 1922.

Clifton, A. R. (Chairman). "Report of Committee on Teacher Rating, Southern Council of Education, California Teachers Association," *Los Angeles School Journal*, 7:31, 33, 37, November 5, 1923.

Collings, Ellsworth. "A Conduct Scale for the Measurement of Teaching," *Journal of Educational Method*, 6:97–103, November, 1926.

Cox, R. M. "Score Cards: Self-Aids for Teachers of Silent Reading in the Intermediate Grades," *Sixth Annual Conference on Elementary Supervision*. Bloomington: Bureau of Coöperative Research, Indiana University, 1929, p. 12–27.

Freyd, Max. "The Graphic Rating Scale," *Journal of Educational Psychology*, 14:83–102, February, 1923.

———. "A Graphic Rating Scale for Teachers," *Journal of Educational Research*, 8:433–39, December, 1923.

Gilmore, M. E. "Judging and Rating the Teacher," *Educational Review*, 74: 269–72, December, 1927.

Kinder, J. S. "A Rating Scale for Practice Teachers," *Education*, 46:108–14, October, 1925.

Koch, H. C. "Practicable Co-operative Supervision," *American School Board Journal*, 68:42–43, March, 1924.

Mead, A. R. "Methods of Studying the Equipment of Teachers Who Do High-Grade Teaching," *Educational Research Bulletin (Ohio State University)*, 5:311–15, 321–23, October 20, 1926.

Mead, C. D. "Scaling Lessons Taught," *Journal of Educational Method*, 6:115–19, 168–74; November, December, 1926.

Pulliam, Roscoe. "Harrisburg Self-Administering Classroom-Activity Test," *Elementary School Journal*, 29:421–27, February, 1929.

Remmers, H. H. and Brandenburg, G. C. "Experimental Data on the Pur-

due Rating Scale for Instructors," *Educational Administration and Supervision*, 13:519–27, November, 1927.

Salm, C. K. "A Score Card for Judging the Recitation," *School Review*, 35:281–85, April, 1927.

Weidemann, C. C. "A New Type Letter of Recommendation for Teachers," *University of Nebraska Publication*, Educational Research Record Vol. 2, No. 2. Lincoln: University of Nebraska, 1929, p. 67–95.

Witham, E. C. "School Measurement," *Journal of Educational Psychology*, 5:571–88, December, 1914.

CHAPTER XVIII

SCORES, NORMS, AND STANDARDS

Introduction.—At the risk of some duplication of material in Chapter III and elsewhere, this chapter combines in one discussion a somewhat detailed treatment of the various types of scores, norms, and standards suggested. Many of those mentioned at least briefly in this chapter have received most if not all of their use in connection with the elementary school, but are included in order that the reader may be prepared to understand the numerous references to them found in educational literature. As will be shown later, comparatively few of them are satisfactory for use in connection with high-school testing.

Point scores.—The expression "point score" refers to a score obtained directly by counting or adding up the points of credit allowed for a pupil's correct answers minus whatever deduction, if any, is made for his incorrect ones. For example, if a pupil answers correctly eighteen direct questions each of which counts one point, his point score is 18, or if a discussion examination contains four questions with values of 20, 20, 20, and 40, respectively, the point score of a pupil who answers the first three entirely correctly is 60. Sometimes "raw score" or "crude score" is used synonymously, but it is better to preserve a distinction between the two. According to this, a raw or crude score is one that has not been modified in any way, but is in its original form, whereas a point score may be one that has been changed in some way, provided it is still expressed in points. To illustrate the difference, suppose that a pupil has made a score of 48 on a certain test known to be twenty points more difficult than some other test. Forty-eight is then his raw, crude, or point score, but if twenty is added to secure his equivalent score on the other test, the result, 68, is a point score, but not a raw or crude score. For high-school tests most norms are reported in terms of point scores.

Derived scores.—A derived or transmuted score is one result-

ing when the original score has been changed or transmuted to some other basis. Thus the score of 68 referred to in the preceding paragraph is such a score. The term is more commonly applied, however, to scores no longer expressed as point scores, but as age, quotient, ratio, or some other variety of scores. Many such scores exist, and those that are commonly employed along with a few that are not are explained in the following pages.

There are perhaps two chief reasons why derived or transmuted scores are desirable. One is the need for some common unit that applies to different tests. If, for example, a pupil makes a point score of 38 on one algebra test and of 54 on another, these facts alone tell us little as to his comparative standing on the two tests unless, as is rarely the case, they happen to be of just the same length and difficulty, and to have identical scoring systems. The second reason is to make scores more readily understandable. If the information just cited with regard to a pupil's scores on two algebra tests were given anyone not familiar with the tests, he would have no idea as to whether the pupil's showing was very poor, average, very good, or something else. If, however, scores are expressed in the same units for all tests, or units such that they indicate what the pupil should do as well as what he does, scores will be much more readily understood by teachers, pupils, parents, and all others interested as well as be directly comparable with one another.

Age scores.—Undoubtedly the most common score of this type is the mental age (M.A.). This was originally employed in connection with the Binet-Simon Scale,[1] but has now come to be used with many, perhaps most, intelligence tests and scales intended for elementary-school children and with quite a number of those for use above this level. A pupil's mental age is simply his score on an intelligence test expressed in terms of age. A mental age of any given amount indicates that average pupils of that chronological age make the point score equivalent to it. Thus if, as is true for a certain test, a point score of 90 is equivalent to a mental age of fourteen years and six months, it means that the average score made by a random or unselected group of pupils whose chronological age is fourteen years and six months is 90. Chiefly for the

[1] Binet, A. et Simon, T. "Le Développement de l'Intelligence chez les Enfants," *L'Année Psychologique*, 14:1–90, 1908.

reason that it is impossible to determine satisfactory mental age equivalents of point scores for children above the middle teens,[2] mental-age scores are not nearly so widely used in the high as in the elementary school. Indeed, some of the intelligence tests intended for high-school use make no provision for transmuting point scores into mental ages. In the case of most others there is no provision for doing so above a mental age of eighteen or nineteen.

The most common age score employed in connection with tests of subject matter is the achievement age (A.A.), also commonly known as the accomplishment age and occasionally as the attainment age. This is similar to the mental age, differing only in the fact that it is derived from an achievement and not an intelligence test. In other words, an achievement age of a given amount, for example thirteen years, corresponds to the average score of all thirteen-year-old individuals. Because of the difficulty of getting together a random or representative group of children of a single chronological age, it is common to determine achievement ages from groups of children of the same mental age. Thus if the average point score on an achievement test made by pupils of mental age thirteen is 62 points, this score will be considered equivalent to an achievement age of thirteen. In the long run whether or not achievement ages are determined from chronological age groups or mental age groups makes no difference.

Some workers in the field have thought that achievement age, which is commonly used to refer either to the age in a single subject or to the average in a group of subjects, is too general an expression and have recommended instead that subject age (S.A.) be applied to a pupil's age in a given subject, and educational age (E.A.) to his average age in a group of subjects. All three terms, achievement age, subject age, and educational age, are now in fairly common use. Also instead of adding the name of the subject, such expressions as "history age," "reading age," and so on are used.

The objections to the use of mental age in high school are equally valid with regard to the other ages just mentioned. There is also the added fact that in the case of most high-school subjects pupils

[2] The chief reason for this is that it is very difficult if not absolutely impossible to secure a random group of pupils above fourteen or fifteen years of age.

have less opportunity to "absorb" the content, or gain a knowledge of it outside formal school work, than is true for most elementary subjects. For example, the pupil who has studied no Latin ordinarily does not pick up any appreciable amount of knowledge of that subject in his ordinary environment, whereas one who has not studied arithmetic would probably acquire some ability in computation outside the school, the amount depending to a considerable extent upon his age, or, in other words, upon how long he had been exposed to his environment. Most workers in the field are agreed, therefore, that age scores should rarely if ever be employed in connection with high-school measurement.

Quotient and ratio scores.—The intelligence quotient (I.Q.) was the first score of this sort to come into use in connection with educational work. It is determined by dividing a pupil's mental age by his chronological age, that is, I.Q. $=\frac{\text{C.A.}}{\text{M.A.}}$. In writing it the decimal point is omitted. Thus if a pupil has a mental age of fifteen and a chronological age of twelve, his I.Q. $=\frac{15}{12}$ or 125. A normal or average I.Q. is, of course, 100, since the mental age of the average pupil must be the same as his chronological age. This same condition holds for all quotient and ratio scores. Since the development of whatever is measured by most general intelligence tests does not appear to proceed regularly past a disputed age, variously set at from fourteen to eighteen years, it is the conventional practice to use a fixed divisor, sixteen being the most common, for all persons whose chronological age is at that point or above. This is not an entirely satisfactory procedure and, coupled with the fact that mental age equivalents are difficult to determine in the upper teens and above, has caused I.Q.'s to be much less commonly employed in high than in elementary school.

Just as the intelligence quotient is computed by comparing intelligence with supposed capacity for intelligence, so achievement has been compared with supposed capacity to achieve as represented by intelligence. The achievement or accomplishment quotient (A.Q.),[8] the measure most frequently employed for this

[8] Monroe, W. S. and Buckingham, B. R. "Illinois Examination: Teacher's Handbook." Urbana: University of Illinois, Bureau of Educational Research, 1920. 31 p.

———, "The Illinois Examination I and II: Teacher's Handbook."

purpose, is obtained by dividing the achievement or accomplishment age by the mental age, that is, A.Q. $= \frac{\text{A.A.}}{\text{M.A.}}$. Although this measure appeared to be coming into rather general use, at least in the elementary-school field, a different suggestion was made before it was thoroughly established. This was that the term "accomplishment ratio" (A.R.) [4] be employed instead of accomplishment or achievement quotient and that A.Q. be defined as $\frac{\text{A.A.}}{\text{C.A.}}$, in other words, compare achievement with chronological age. At the present time usage is divided, but the writer believes that the majority of workers in the field are following the original suggestion and not employing ratio scores at all.

In addition to those just mentioned, both subject quotients (S.Q.) and educational quotients (E.Q.), subject ratios (S.R.) and educational ratios (E.R.), especially the former, are receiving some use. These are used without any confusion of meaning. The first pair involve the comparison of subject and educational age, respectively, with chronological age, that is, S.Q. $= \frac{\text{S.A.}}{\text{C.A.}}$ and E.Q. $= \frac{\text{E.A.}}{\text{C.A.}}$. The latter pair involved a comparison with mental age, thus S.R. $= \frac{\text{S.A.}}{\text{M.A.}}$ and E.R. $= \frac{\text{E.A.}}{\text{M.A.}}$. That the quotient and ratio scores for the comparison of achievement with either mental or chronological age are inappropriate in the high school necessarily follows from the fact that the age scores upon which they are based have no place there.

Bloomington, Illinois: Public School Publishing Company, 1920. 32 p.

Monroe, W. S. "The Illinois Examination," *University of Illinois Bulletin*, Vol. 19, No. 9, Bureau of Educational Research Bulletin No. 6. Urbana: University of Illinois, 1921. 70 p.

Franzen, R. H. "The Accomplishment Quotient of School Marks in Terms of Individual Capacity," *Teachers College Record*, 21:432–40, November, 1920.

[4] *Ibid.* "The Accomplishment Ratio," *Teachers College, Columbia University. Contributions to Education*, No. 125. New York: Bureau of Publications, Teachers College, Columbia University, 1922. 59 p.

———— "The Conservation of Talent," Terman, L. M. et al *Intelligence Tests and School Reorganization*. Yonkers, New York: World Book Company, 1922, Chapter IV.

Several other derived scores similar to quotient and ratio scores, but not having either terms in their names, have also been suggested. Among these are the coefficient of brightness,[5] the index of brightness,[6] and the coefficient of intelligence,[7] and the index of intelligence. As their names imply, all four have to do with intelligence test scores. They are similar in principle to the intelligence quotient in that they involve the comparison of an individual's intelligence score with the normal score for one of his age, but differ in the way in which this comparison is made. All are 100 for average individuals, as is the I.Q., also above this for those of superior and below it for those of inferior intelligence, but as their values differ more from 100, either up or down, they tend to differ more from the I.Q. None of them has received any very wide use, even though they seem to possess certain theoretical advantages over intelligence quotients in the case of individuals above the middle teens at least.

T-[8], C-[9], and B-scores.—One of the best-known derived scores is the *T*-score, so named by McCall in honor of Terman and Thorndike. It is a score given according to the *T*-scale, which is based upon the distribution of ability of a random or complete group of twelve-year-old pupils. The scale consists of one hundred units of one tenth standard deviation ($.1\sigma$) each, and extends from five standard deviations below the mean of twelve-year-old pupil ability to five above it. Therefore an average twelve-year-old score is 50. For pupils whose abilities do not differ too much from those of twelve-year-olds, this provides a fairly satisfactory comparable score, but many high-school pupils are too far above twelve-year-old ability to be satisfactorily measured by it. Despite this fact a number of high-school tests in common with many for the ele-

[5] Otis, A. S. *Statistical Method in Educational Measurement.* Yonkers, New York: World Book Company, 1925, p. 153–55.

[6] *Ibid.*, p. 155–56.

Freeman, F. N. *Mental Tests.* Boston: Houghton Mifflin Company, 1926, p. 283–84.

[7] *Ibid.*, p. 134, 281–82.

[8] McCall, W. A. "Scaling the Test—*T* Scale," *How to Measure in Education.* New York: The Macmillan Company, 1922, Chapter X.

[9] Van Wagenen, M. J. *A Teachers' Manual in the Use of the Educational Scales.* Bloomington, Illinois: Public School Publishing Company, 1928, p. 6–21.

mentary school provide transmution tables by which point scores may be changed into *T*-scores.

C-scores and the *C*-scale suggested by Van Wagenen are in many ways similar to *T*-scores and the *T*-scale. The chief difference is that the unit used is one-tenth quartile deviation (.1Q) instead of one-tenth standard deviation. The scale extends the same distance, from -5σ to $+5\sigma$, a distance equal to 14.8 quartile deviations; therefore the scale has 148 units instead of the 100 of the *T*-scale. Although *C*-scores are provided for several high-school tests, they also are not generally employed at that level.

B-scores, so named in honor of Binet and Buckingham, are the same as grade scores. They are composed of two figures, one in units' place and one in tenths' place. The first indicates the grade and the second the month of the school year. Thus, for example, a *B*-score of 7.8 indicates that the pupil receiving it is doing work that is just average for pupils in the eighth month of the seventh grade. Such scores are almost never employed in high school. The chief reason is that there is such great difference of practice as to the year or grade in which subjects are offered, and therefore that the mere fact that a pupil is in a given grade is, in high school, not a safe indication as to which his achievement should be.

Index of effort or studiousness.—Several other derived scores of various sorts have been suggested, but they either have so little to recommend them or have received so little use, or both, that it does not seem worth while even to name them here. The one exception to this general statement, in the opinion of the writer, is the index of effort or of studiousness suggested by Symonds.[10] He employed the term in a rather general way to include various more or less similar methods of comparing achievement with capacity and suggested two methods in particular. The first and simpler of these consists merely in ranking pupils according to their capacity and achievement and then taking the differences between the ranks. A difference in favor of achievement corresponds, of course, to an achievement quotient or ratio of more than 100, and one in favor of intelligence to an achievement quotient or ratio of less than 100. His second method is somewhat more difficult. It requires that both achievement and general intelligence test scores be

[10] Symonds, P. M. *Measurement in Secondary Education*. New York: The Macmillan Company, 1927, p. 521–25.

turned into standard deviation units, that is, into distances from the average score expressed in terms of the standard deviation. The difference between a pupil's achievement and his intelligence when both are expressed in standard deviation units indicates whether or not he is doing work above or below that of pupils in general.[11]

Of these two methods the former is probably to be preferred. The chief reason for this is its simplicity. The latter requires more work than many teachers are willing to undertake, and it is not apparent that the results from its use are sufficiently more valid or reliable than those from the other to justify the additional labor. Both suffer from certain weaknesses, the chief one being that they may be used only for purposes of comparing pupils within the same group and not those of one group with those of another.

Varieties of norms and standards.—From the standpoint of the type of scores in which they are expressed, there are several commonly used varieties of norms and standards. The most frequently employed is probably the grade norm or standard, that is, the statement of what pupils in given grades actually do or should do. Age norms and standards are somewhat less frequently employed, but still are not unusual. The same is true of mental age norms and standards, which are essentially the same as ordinary age ones, but are usually determined in a somewhat different manner. As has been stated in connection with the discussion of these various types of scores, none of the three is appropriate for much if any use in high school. Instead norms or standards on the basis of time studied are the best now employed in any considerable number of instances. These show what pupils actually do or should do at the end of one semester, two semesters, three semesters, and so on, or sometimes even at the ends of shorter periods of time. Occasionally they are only given by years, but it is becoming increasingly common for the semester to be the unit. Even these as usually given are subject to at least one serious criticism. This is that because of both added maturity and knowledge pupils who have studied a given subject a certain length of time during a

[11] In order to avoid fractions and to eliminate negative numbers, Symonds suggests that the actual difference be multiplied by 10 and added algebraically to 50.

later portion of their high-school course ordinarily do better work therein than do those who have taken it earlier. For example, juniors beginning Spanish, especially if they have had Latin or French during the first two years, usually do more and better work in one semester of study than do freshmen who have had no foreign language previously. This is at present very rarely recognized in the manner of giving norms, but it is to be hoped that norms on this basis will become increasingly common. In the case of a few subjects or portions of subjects practice is nearly enough in agreement as to the year in which they are taken that it is not highly important, but in the case of half or more of the high-school subjects this is not the case.

The basis of norms and standards.—The previous discussion has been designed to make the point that in reporting scores and in setting up norms and standards in connection with high-school measurement, it is generally best to use point scores and to state norms or standards in terms of the length of time subjects have been studied. These, however, are not all the factors that have to do with determining satisfactory norms and standards. Probably the first additional question to arise is that of what scores or points in the general distribution of scores are employed as norms. As has been stated, the median is the most commonly reported point, next to that the first and third quartiles, and next certain percentile points, especially the fifth, tenth, twenty-fifth, fiftieth, seventy-fifth, ninetieth, and ninty-fifth. Sometimes, but not commonly, standards as well as norms are determined solely from the scores actually made by pupils. In some cases the third quartile has been set as the standard on the assumption that if one-fourth of all pupils can do better than that it is not unreasonable to expect the whole group to come up to it. In some cases even higher points, such as the ninetieth percentile, have been taken as standards. Occasionally several different points have been taken as standards for as many groups of pupils. Perhaps the ninetieth or even the ninety-fifth percentile is set up as the standard for superior pupils, the third quartile for average pupils, and the median or even the first quartile for inferior pupils, thus holding before each group a standard that represents practically the upper limit being achieved by any of that group. In most cases, however, standards are not determined directly from actual

scores, although the scores made by pupils play a part in setting them.

It has already been stated in Chapter III that it is desirable to supplement general nationwide norms with state norms, city norms, norms for different types of school organization, and so forth. There are, indeed, many factors that cause norms to vary. In any particular class or school any of these that are present should be taken into consideration in the use of norms. The achievements of pupils vary not only according to locality, race, sex, general plan of school organization, and so forth, but also because of such factors as methods of teaching, textbooks employed, length of school term or year, incentives and motivation provided, conditions under which tests are given, including previous practice with similar tests or coaching upon such tests, whether pupils are required to take the subject or have elected it, and so on. In most instances it is impossible to make exact or even approximate numerical allowances for the effects of these causes, but in many cases one can be reasonably sure whether they result in higher or lower scores than would otherwise be the case. Other things being equal, it would ordinarily be true, for example, that pupils from a ten-months school make better scores than those from a nine-months school, and that pupils who have elected a subject do better work than if they have been required to take it. The situation may be summed up somewhat as follows: if the scores made by a group of pupils tend to differ very much in either direction from general norms, it should not be considered positive proof of superior or inferior work on the part of the pupils, but rather an attempt should be made to determine the causes and then to interpret the results in the light thereof.

There are also many factors that enter into the setting of satisfactory standards of achievement. The amount of knowledge or ability needed to meet vocational demands and others from outside the school or those of subjects to be carried later, the relative values of the outcomes of certain subjects as compared with those of others, what pupils show that they can accomplish with a reasonable expenditure of time and effort and a fairly good school environment, are among the most important factors to be considered. In some subjects, such as spelling and arithmetic, it is commonly agreed that perfect accuracy should be set as the standard,

in others there is considerable difference of opinion as to how high a degree of accuracy should be required. For example, how accurate measurements should a pupil learn to make in physics or chemistry laboratory work? Not only accuracy or quality but speed or rate standards must also be set in at least some subjects. Frequently the two are very closely connected. In mathematics and typing, for example, increase of speed is liable to result in decrease of accuracy so that standards must be set that represent a desirable compromise between these two qualities.

It is unfortunately true that most authors and publishers of texts have not attempted to set up standards, but have merely reported norms and left the determination of standards to those using the tests. A few workers in the field have in some way or other determined what seem to them appropriate standards, commonly for only one or a very few tests. Among the outstanding attempts to do this is that of Symonds,[12] who has, on the basis of a considerable amount of experimental testing, set up standards for nine widely used high-school tests. They are what in his opinion pupils of various degrees of intelligence according to the Terman Group Test of Mental Ability should do on each test December first, March first, and June first.

Comparing scores with one another.—There is need for considerable caution in comparing scores made by different pupils or by the same pupil at different times upon the same test. For example, let us suppose that on a given test Pupil *A* makes a score of 40, Pupil *B* of 50, and Pupil *C* of 60. Although it might possibly be otherwise,[13] it is ordinarily safe to interpret these scores as indicating that at the time he took the test *B* knew more about the items dealt with than did *A*, and likewise that *C* knew more than either *B* or *A*. This is, of course, on the assumption that their scores were honestly made and that they did not purposely respond correctly to fewer items than they knew or were able to

[12] Symonds, P. M. *Ability Standards for Standardized Achievement Tests in the High School.* New York: Bureau of Publications, Teachers College, Columbia University, 1927. 91 p.

[13] If for some reason *A* responded to forty relatively difficult or complex questions or exercises whereas *B* responded correctly to fifty much easier or simpler ones, it might be that *A*'s score of 40 would indicate more knowledge than *B*'s of 50. This happens rarely enough, however, that it need receive little consideration.

do in the time allowed. It is not at all safe, however, to assume that *C*'s knowledge exceeds *B*'s by the same amount as *B*'s exceeds *A*'s, or, in other words, that equal increments of score represent equal increments of knowledge or ability. For example, if the test contained forty decidedly easy elements, ten somewhat more difficult, and ten or more very much more difficult than either of the other groups, and if *A* answered only the first forty, *B* those plus the ten somewhat more difficult, and *C* all these plus ten of the very difficult ones, the excess of *C*'s knowledge over *B*'s must be much greater than that of *B*'s over *A*'s. It is only in the case of uniform tests containing elements of approximately the same degree of difficulty that it can be assumed that equal differences in scores represent equal differences in knowledge or ability. If a test is scaled, and ordinarily also if it is irregular with exercises of considerable difference in difficulty, the higher a pupil's score the more added ability is required to raise it a certain number of points. In other words, on such a test it requires more ability to increase a score of 40 to 50 than one of 30 to 40, still more to raise one of 50 to 60, and so on. This is merely another way of saying that equal differences in scores represent progressively large differences in actual ability the larger the scores concerned are.

The comparison of scores by dividing one into another and thus determining what proportion or ratio one is of another is usually even less justifiable than the procedure referred to in the last paragraph. There are two chief reasons why this is true. The first is the same as that just presented, that the units or score points employed are not the same all through the test, so that a score of 40, for example, does not represent twice as much ability as one of 20, but probably three or four or even more times as much. As before, this objection does not hold in the case of tests wherein the elements are of approximately equal difficulty.

The other reason, however, is one that exists to a greater or lesser degree in the case of most tests intended for high-school use. It is that the zero point of the test is not the true zero point of the ability or trait measured. In other words, since the easiest element in the test is not the easiest possible one in the subject matter covered, a score of zero on the test does not represent an absolute lack of whatever is being measured. Instead a score of zero ordinarily means merely that the amount of what is being

measured is relatively small but undetermined. Similarly a score of any given amount does not mean that many points above a true zero point, but that many points above the zero point of the particular test. The situation is similar to one in which the heights of individuals are being measured by a stick that contains a portion of unknown length not marked off into linear units and another portion that is known and marked. Using such a measuring instrument it might be found that one person's height was the portion of unknown length plus three feet five inches; another person's was the portion plus three feet seven inches; that of another the portion plus three feet ten inches, and so on. From such measures we can tell the differences in heights of the different individuals, that the second is two inches higher than the first, and the third three inches taller than the second, but we cannot tell what proportion the height of the first is of that of the second, and so on. In other words, the height of the first may be written as $x+3$ feet 5 inches, that of the second as $x+3$ feet 7 inches, and that of the third as $x+3$ feet 10 inches. Similarly if a pupil makes a score of 20 on a test, his score as measured from an absolute zero point is $x+20$, if another pupil makes one of 40, his is $x+40$, and so on. Thus even if the units are equal it is readily seen that the second pupil's score is not double that of the first.

There are a few tests in which the zero point on the test approaches very near to the absolute zero point. In Thorndike's Handwriting Scale, for example, zero was taken as the value of a specimen that, according to the combined opinions of a number of judges, could be recognized as attempted handwriting, but not read. Similarly for his drawing scale Thorndike took zero as the value of a specimen recognized as an attempt to draw, but not as a drawing of the object it was intended to represent. In subjects in which the desired outcomes of achievement are not material in the sense that writing and drawing are, this method cannot be applied. It is, however, possible in some subjects to choose an item of information or skill so easy that any person who does not know it or is unable to do it has approximately zero knowledge or ability. For example, in arithmetic the person who cannot add $1+1$ approaches zero ability very closely. In solving algebraic equations the person who cannot solve such an equation of $x+1=2$ like-

wise has practically zero ability. In Latin, anyone who does not know that the word *"et"* means "and," must have very nearly zero ability. In cooking, the pupil who is not able to bring a pan-ful of water to the boiling point approximates zero ability. In such a subject as literature, however, it is very difficult to pick out any item of knowledge that is easier than, or at least as easy as, any other, and the lack of which indicates that a pupil has no knowledge of the subject. In American history one of the commonest items of knowledge is that the Declaration of Independence was signed in 1776, but an individual may know many facts about American history without happening to know this particular date. Therefore even if the attempt were made, it would in many cases be impossible to make the zero point of particular tests absolute zero points in the subject matter dealt with. In comparatively few cases, however, has this been attempted, so that the actual zero points of tests range all the way from approximately the true zero point up to points considerable but unknown distances above it, some of them at least so far above that after studying a subject several weeks or even months pupils may still make zero scores.

Summary.—After defining point scores and derived scores, this chapter takes up a number of scores of the latter type. These include the mental age, achievement, accomplishment or attainment age, subject age, educational age, intelligence quotient, achievement or accomplishment quotient, achievement or accomplishment ratio, subject quotient, educational quotient, subject ratio, educational ratio, coefficient and index of brightness and also of intelligence, *T*-, *C*-, and *B*-scores, and index of effort or studiousness. It is shown that point scores and the index of effort are the only ones appropriate for general use in high school. A number of the others are valuable in the elementary school and perhaps occasionally in high school. Grade, age, and mental-age norms and standards are also shown to have little place in high school, whereas those based on length of time a subject has been studied are suitable. Several factors that need to be taken into consideration in setting and interpreting norms and standards are also mentioned and discussed. Finally it is pointed out that in comparing different scores on the same test a considerable degree

of caution is necessary, since the same differences occurring at different places on the scale do not indicate equal differences in ability, and also since scores on most tests cannot be divided into one another with valid results.

BIBLIOGRAPHY

Foran, T. G. "The Meaning and Limitations of Scores, Norms, and Standards in Educational Measurement," *Catholic University of America Educational Research Bulletin*, Vol. 3, No. 2. Washington: Catholic Education Press, 1928, p. 16–19, 23–26.

Greene, H. A. and Jorgensen, A. N. "Interpreting the Results of Testing," *The Use and Interpretation of Educational Tests*. New York: Longmans, Green and Company, 1929, Chapter VII.

Kelley, T. L. *Interpretation of Educational Measurements*. Yonkers, New York: World Book Company, 1927, p. 5–9, 22–25, 34–35.

———. "A New Method for Determining the Significance of Differences in Intelligence and Achievement Scores," *Journal of Educational Psychology*, 14:326, September, 1923.

Monroe, W. S. *An Introduction to the Theory of Educational Measurements*. Boston: Houghton Mifflin Company, 1923, Chapters V, VI, VII, and VIII.

Monroe, W. S., DeVoss, J. C., and Kelly, F. J. "The Meaning of Scores," *Educational Tests and Measurements*, Revised and Enlarged Edition. Boston: Houghton Mifflin Company, 1924, Chapter XII.

Nygaard, P. H. "A Revised Accomplishment Quotient," *Journal of Educational Research*, 18:87, June, 1928.

Odell, C. W. "A Critical Study of Measures of Achievement Relative to Capacity," *University of Illinois Bulletin*, Vol. 26, No. 29, Bureau of Educational Research Bulletin No. 45. Urbana: University of Illinois, 1929. 58 p.

Peters, C. C. "A Method for Computing Accomplishment Quotients on the High-School and College Levels," *Journal of Educational Research*, 14: 99–111, September, 1926.

Popenoe, Herbert. "A Report of Certain Significant Deficiencies of the Accomplishment Quotient," *Journal of Educational Research*, 16:40–47, June, 1927.

Rand, Gertrude. "A Discussion of the Quotient Method of Specifying Test Results," *Journal of Educational Psychology*, 16:599–618, December, 1925.

Ruch, G. M. "The Achievement Quotient Technique," *Journal of Educational Psychology*, 14:334–43, September, 1923.

Sherrod, C. C. "The Development of the Idea of Quotients in Education," *Peabody Journal of Education*, 1:44–49, July, 1923.

Symonds, P. M. *Measurement in Secondary Education*. New York: The Macmillan Company, 1927, Chapters XIII and XV.

Toops, H. A. and Symonds, P. M. "What Shall We Expect of the A.Q.?"

Journal of Educational Psychology, 13:513–28, December, 1922; 14:27–38, January, 1923.

Torgerson, T. L. "The Efficiency Quotient as a Measure of Achievement," *Journal of Educational Research*, 6:25–32, June, 1922.

Wilson, W. R. "The Misleading Accomplishment Quotient," *Journal of Educational Research*, 17:1–10, January, 1928.

CHAPTER XIX

SCHOOL MARKS

Should school marks be employed?—The question has at times been raised whether it would not be better to abolish school marks of the ordinary sort, or at least to employ only two marks, such as passing and failing, or satisfactory and unsatisfactory. It has been urged that marks become too largely the chief incentives motivating pupils and thus lead them to work chiefly for ulterior and artificial rewards rather than through interest in the subjects carried or other more satisfactory motives. Furthermore, it is claimed that the usual marking system produces such undesirable results as overwork, cheating, feelings of self-conceit on the part of some and of discouragement on that of others, jealousy, and so forth. The general answer that may be made to these objections is that most of the outcomes mentioned may result from unsatisfactory marking systems or overemphasis upon marks, but are not necessary consequences of marks in themselves. Furthermore, most persons are accustomed to formal or informal ratings of their abilities, accomplishments and so forth outside of school, so why avoid them in school?

There are also positive arguments in favor of marks both from the standpoint of the pupil and from that of the school. It appears to be a trait in human nature to do better work if what is done is rated in some satisfactory way, and school marks fill this need in regard to school work. Moreover in education as in business and industry it is important from the standpoint of administration and organization that the products, in this case the achievements of pupils, be evaluated. It is generally considered that the more accurately products are rated the more likely is the concern producing them to succeed, since those directing it know more about its operations and what, if any, changes are needed. There appears to be no reason why this does not apply in school work as well as elsewhere. There is a real need for marks or some more or less

similar means of reporting on the achievements of pupils to themselves, to their parents or guardians, to prospective employers, to other educational institutions, and to any others who are interested. For these and other minor reasons the writer, apparently in common with the vast majority of educational workers, believes that school marks should be employed. If this is granted, it follows that an effort should be made to make them as valid and reliable as possible.

The basis of marks.—Probably the chief weakness in school marks as actually employed is that the bases on which they are assigned by different teachers and schools vary so greatly. Among the things upon which they are based are any one of the following or any combination thereof: achievement, attitude, interest, effort, improvement, quality, quantity, initiative, and so on. Sometimes these are determined from a final examination alone, sometimes from daily oral work, from written work prepared outside of class, from laboratory work, or from combinations of these and other phases of pupil activity in various proportions. Even among teachers who are using the same bases there is no assurance that the same marks mean the same things. One may, for example, give a barely passing mark to a pupil who has mastered certain minimum essentials, another to a pupil who has done 70 per cent as much work as the best pupil in the class, or 70 per cent of all the assigned work, another to one who seems just able to carry more advanced work in the same subject, and so on. Likewise the highest possible mark, whatever it may be, is sometimes interpreted as meaning absolute perfection, sometimes as signifying merely the best work in the class, sometimes that all the work assigned has been done satisfactorily, sometimes that a pupil has done all he can reasonably be expected to do, and so on.

Because of this great variety in the meanings of marks given by different teachers there is need that measures be taken to produce a greater degree of uniformity. At present it can hardly be hoped that large groups of schools or school systems will unite in adopting marking systems sufficiently uniform to be satisfactory, even though some progress has already been made in this direction and more is being made. It is, however, a reasonable expectation that in every school or school system a uniform scheme be adopted. Uniformity should include not only the use of the same marks,

passing point, and so forth, but should also extend to the definitions of the significance of these marks. In addition to such general specifications for a whole system or school, each department in high school should go further and draw up certain supplementary specifications, appropriate to the subject matter dealt with, for use by the teachers therein. For example, mathematics teachers should agree as to whether to require complete accuracy for any credit or to give partial credit if a response is partially correct; composition teachers should agree as to what to do about factual errors in pupils' themes, and so on.

It is desirable that the specifications of a marking system should grow out of a coöperative study of the matter by all the teachers concerned rather than that they be made out and handed down by some one or a few higher officials of the system. If the number of teachers concerned is very large, it will, of course, be necessary to have much of the work done by one or more committees, but it should be handled in such a way that the teachers feel it is their marking system. In many cases it will be desirable, if not absolutely necessary, to lead the teachers to feel the need of a new marking system before the actual construction of one is begun. Probably the best method of starting the movement is to have different teachers score the same papers and then reveal and discuss the differences in the marks given. Most teachers will agree that it is undesirable that the mere fact of whether a pupil happens to be under one teacher or another should make a large difference in the mark he receives, and will be willing to consider how this condition may be corrected.

A number of sets of marking standards or specifications varying from rather simple ones to decidedly elaborate schemes have appeared in print, and many others have been worked out by groups of teachers for their own use but never published. As examples of the former, the reader is referred to the one employed in the Beechview-Beechwood Public Schools[1] and also to that used at Geneseo.[2] Each defines five qualities of work for each of several points. Those of the first are knowledge of subject matter, prep-

[1] Masters, H. G. "Standards for Rating Pupils," *Journal of Educational Method*, 1:176–77, January, 1922.

[2] Reeder, J. C. "The Geneseo Scale of Qualities," *Elementary School Journal*, 20:292–96, December, 1919.

aration, attitude and application, whereas the second omits application. There are subdivisions into such points as recitation, questioning, degree of accuracy and so on. As an example, the specifications as to the asking of questions in the first set referred to are given below. They are arranged in order from best to worst:

Asks intelligent questions
Asks majority of questions intelligently
Asks some intelligent questions
Seldom asks intelligent questions
Asks and answers very few questions

Other more or less similar scales rely more largely on adjectives of quality or degree than on fairly definite statements of pupil activity. In general these are less satisfactory since there is less likelihood that they will mean the same to different persons.

There is more or less agreement that if only a single mark is to be given in each subject it should be as nearly as possible a mark of absolute achievement and not involve directly such factors as intelligence, interest, attitude, effort, and so forth. There is no doubt that it is very helpful in dealing with pupils to have marks or ratings of these other traits and abilities. Effort should be recognized and neither the pupil of inferior intelligence allowed to become discouraged when he is working somewhere near his maximum, even though his absolute achievement is low, nor should the superior pupil be allowed to feel satisfied because his mark is somewhat above average when he is capable of doing much better. Pupils whose attitude toward their work is favorable should be encouraged therein and those whose attitude is unfavorable should be so dealt with as to modify it. Therefore, it is recommended that a supplementary marking or rating system that takes account of some of the most important of these other factors be employed. Comparatively few high schools do this, chiefly because of the additional labor involved, but if the ratings are carefully made and wisely used they will much more than justify the expenditure of time and energy required.

As already suggested, one of the questions that arises in connection with the determination of semester or other final marks is that of how much to count on each of the several portions or phases of the pupil's work. Probably the chief principle that can be laid down with respect to this is that general hard and fast rules

are unsatisfactory. The proper proportions differ according to subjects or even portions of subjects, according to points of emphases, according to methods of conducting classes, and so forth. In some cases it is possible to test the work of a semester much more thoroughly by a single final examination than in others. Thus a single good translation examination in foreign language comes much nearer to testing the whole semester's achievement or mental growth in that subject than any written examination in manual training or cooking, or even in such a subject as history or physics. In some courses papers, notebooks, and other written work have a much larger place than in others and should, therefore, play a larger part in determining final marks. The writer believes that on the whole the tendency is to count too much upon the final and perhaps other fairly long examinations and too little upon oral daily class work and short tests. As a general rule to which, however, there might be occasional justifiable exceptions, he recommends that the final examination should not count for more than 25 per cent of the final mark, that all formal written examinations and tests combined should not count for more than 50 per cent, and that daily class work including oral work, laboratory work, outside written work and so forth should count for at least 50 per cent. Usually written examinations should count for less in such subjects as manual training, home economics, physical education, and so on, in which there is comparatively little book work, than in the ordinary academic subjects.

What marks should be used?—The diversity in the marking symbols employed is as great as that in the bases upon which marks are given. For example, a study made by the writer a few years ago [8] revealed the fact that among less than three hundred Illinois high schools responding to a questionnaire there were, if all minor variations be counted, about one hundred different systems of marking in use. Even when such minor variations as did not constitute material differences were eliminated, there still remained about thirty different systems, differing either in the fact that they varied markedly in the symbols employed or in the meanings of the symbols. Studies made by others show that the situation found in Illinois is not exceptional, but instead typical. Evidently,

[8] Odell, C. W. "High School Marking Systems," *School Review*, 33:346–54, May, 1925.

therefore, there is little consensus of opinion concerning the matter.

In general all marking systems may be grouped into two classes, those that employ percentile marks and those that do not. Although the former are probably still more common in high school than are the latter, the writer believes that the latter are to be preferred. Experimental evidence seems to indicate that even our better teachers cannot distinguish more than a comparatively few degrees of ability or difference in achievement, certainly not degrees as fine as are represented by differences of one or two per cent in the ordinary percentile system and probable no finer than are represented by five per cent. Such marks, therefore, tend to produce a false impression of accuracy and to indicate differences where it is doubtful if they exist. Among the undesirable results caused by this are feelings of undue elation on the part of some pupils and of undue discouragement on that of others, the awarding of school honors according to unreliable distinctions in achievement, and so on. For these and other minor reasons a marking system of five or six letters or other symbols seems much preferable to a percentile system of the usual type. The symbols may be either the first few letters of the alphabet, letters that stand for descriptive terms such as *E* for "excellent," *G* for "good," *M* for "medium," *P* for "poor" and *VP* for "very poor," numbers such as 1, 2, 3, 4, 5, or even percentile marks with intervals of five or more. The exact symbols used are of little importance; their meanings, however, are significant. Of these five or six symbols two should be failing marks. The reason for this is that pupils who come close to passing and can probably pass if they repeat the work may be distinguished from those who fall far below and probably cannot pass by repeating unless their failures were caused by lack of application rather than of capacity. The common practice of using plus and minus signs with the symbols does not appear desirable, since its effect is really to change a five or six symbol system into one of fifteen or eighteen. They may well be used occasionally in the case of such a mark as *A* + for exceptionally good work, or as *C* — for that just at passing, but should be decidedly rare. Furthermore, there should be no marks of "conditioned" or anything corresponding to that. In the case of those who have for some good cause been absent from school and

missed a portion of the work or for some other reason seem not able to go ahead and yet do not deserve to be failed, marks should be withheld until they have had an opportunity of doing the work. A pupil's mark for one semester, however, should never depend on that for another.

The application of the normal frequency curve to marking.—Among the questions connected with marking most frequently raised and most heatedly argued during the past few years is that of how far, if at all, marks should conform to the normal frequency distribution.[4] Because it has been found that measurements of most human traits and abilities derived from random or unselected groups conform rather closely to the normal curve, many persons believe and argue that school marks should do likewise. They point out that teachers' marks are usually decidedly subjective and unreliable, that those given by different teachers or by the same teacher at different times do not have the same meaning, and state that following the normal distribution curve is the best means of remedying the situation.

The argument most commonly advanced against this procedure is that although the marks of large enough groups of pupils should perhaps follow this curve, the relatively small groups in ordinary classes are likely to deviate considerably from average and, therefore, that their marks should likewise deviate. There is no doubt that there is some truth in this argument. Indeed, most of those who favor applying the normal curve to the distribution of marks do not believe that it should be applied closely and rigidly in the case of groups of pupils as small as the average high-school class, but rather that it should be followed approximately by a teacher in the distribution of marks to all her pupils, and applied only very loosely to those of single classes. In other words, it should be a general guide rather than an exact rule. Those who believe that it should be used claim, however, and probably truly, that teachers tend to exaggerate differences between classes, to think that some classes are more superior than they are, and others more

[4] It has sometimes been suggested that they should conform to certain other distributions, but these suggestions have been so rare in comparison with those urging the use of a normal distribution that they will not be considered here.

For illustrations and explanations of the normal and several other types of curves, see Chapter XXV.

inferior, and that for this reason more or less approximate adherence to such a curve is desirable in the case of any class containing more than a very few, perhaps a dozen or so, pupils.

Another argument advanced against the plan is that it is too mechanical, that it robs teachers of one of their functions. The answer to this is that practically no one contends that it should be applied in absolutely mechanical fashion without judgment entering in, and that however it is applied the teacher must decide which particular pupils receive each mark.

An additional argument advanced is that because of the effect of various more or less artificial incentives upon school work the resulting distributions of marks cannot be expected to be normal, but will be skew or irregular, with marks bunched at certain places, such as the passing point, the exemption point, the point necessary for inclusion on the honor roll, and scarce between these points. There is no doubt that such incentives do operate to cause something of the effect claimed, but in many cases their apparent effect is due more to their influence upon teachers than upon pupils. Teachers frequently tend to give no pupils very high failing marks, but rather the lowest possible passing mark. Likewise, although probably less frequently, they give this same mark to some pupils who really deserve better ones because they believe these pupils are loafing and need stimulation, do not have the proper attitude, are disciplinary problems or something else. The same is true of other such points as were mentioned, that teachers frequently hesitate to give marks immediately below them and also reduce higher ones to them.

Still another point raised is that because of the difference in subjects, courses, or teachers, or pupils' opinions of them, certain classes tend to draw superior pupils and others inferior ones. Furthermore it is claimed that because of differences in the abilities of teachers, in sizes of classes, in equipment provided, in interest of pupils, and so on, some classes should be expected to do better work than average and should, therefore, receive better marks, and others do poorer work and receive lower marks. Likewise it is urged that as pupils advance further in school, the inferior are more and more eliminated and those who remain should not be marked so low as the whole group at the beginning. Most of these arguments have a degree of truth in them, but do

not really conflict with the application of the normal frequency distribution to marks if it is understood that it is to be followed only approximately and reasonably rather than slavishly; moreover, it is probably that differences due to subject matter or teachers, to size of class or amount of equipment, and to most other alleged causes, are not usually very great. Those in the average ability of pupils at different school levels are undoubtedly real, although not very large within the four years of high school, and should receive some recognition in the assignment of marks.

In view of the arguments and facts presented above and others, the writer recommends that the normal distribution should be taken as a general standard or ideal distribution to be applied fairly closely to most large groups of pupils and very loosely to groups of ordinary class size. When it is definitely known that a group of pupils is superior or inferior in capacity or is doing superior or inferior work, regardless of the cause, he believes some allowance should be made for this in the marks given the members of the group. This allowance should usually not be as great as the variation of the group from average, but perhaps about half as great.

Since there are many ways of applying the normal curve, the mere fact that marks are to follow it does not define exactly how they are to be distributed. Theoretically the curve does not meet its base line except at infinity, so that it must be arbitrarily cut off at the ends.[5] The point at which it is cut off and the number of divisions into which it is divided determine the per cents in each. A number of the more frequently mentioned distributions are given in Table IV.

TABLE IV. SUGGESTED PERCENTILE DISTRIBUTIONS OF MARKS

Part I. Three-Symbol Systems

*A**	*B*	*C*
16	68	16
20	60	20
25	50	25

* In this table as well as in the discussion in the text the first letters of the alphabet are used as the marking symbols without any intention of suggesting that they are better than any other set.

[5] The most usual points at which the normal curve is cut off are 2.5, 3.0,

Part II. Four-Symbol Systems

A	*B*	*C*	*D*
10	40	40	10
15	35	35	15

Part III. Five-Symbol Systems

A	*B*	*C*	*D*	*E*
2	23	50	23	2
3	22	50	22	3
3½	24	45	24	3½
5	20	50	20	5
7	24	38	24	7
10	15	50	15	10
10	20	40	20	10

Part IV. Six-Symbol Systems

A	*B*	*C*	*D*	*E*	*F*
2	14	34	34	14	2
5	15	30	30	15	5

Part V. Seven-Symbol Systems

A	*B*	*C*	*D*	*E*	*F*	*G*
1	6	24	38	24	6	1
3	10	22	30	22	10	3
5	10	20	30	20	10	5

This table includes five parts in which are suggested distributions for systems employing from three up to seven symbols. In those with three, four, or five symbols it is ordinarily assumed that only the last one represents failure, whereas in the six and seven symbol systems usually the last two are failing ones. It includes not only distributions secured by cutting off various portions of the normal curve and dividing the remainder, but also some that are symmetrical but not normal.

Instead, however, of merely adopting a single fixed distribution as the standard, even though it is not to be adhered to closely, the writer believes it is better to adopt limits within which distributions of marks should fall. Although he believes that the best five-symbol system is probably that in which the per cents are 10, 20, 40, 20, and 10, or perhaps the one with 7, 24, 38, 24, and 7 per cent, he would suggest as better than either some such limits as the following: *A*'s 5 to 15 per cent, *B*'s 15 to 30 per cent, *C*'s 25 to 50 per cent, *D*'s 15 to 30 per cent, and *E*'s 5 to 15 per

3.5 and 4.0 standard deviations, or 4.0 or 5.0 median deviations, from the mean. These various distances include from about 98.74 up to about 99.99 per cent of the area under the normal curve, thus neglecting from 1.26 down to .01 per cent of it.

cent. As was stated previously, however, he believes it is better to make use of two failing marks, so that he would either use a five-symbol system with only three passing marks or six symbols four of which are passing. For the former he would suggest some such limits as: *A*'s 5 to 20 per cent, *B*'s 25 to 50 per cent, *C*'s 25 to 50 per cent, *D*'s and *E*'s combined 5 to 20 per cent. For a six-symbol system he would recommend about the same as just given for the first five-symbol system except that the last per cent given for it should include both failing marks. Even with these limits instead of single fixed distributions, however, he would not insist that the marks of every class should fall within them. In many small classes there may be no pupils who deserve *A*'s or none who deserve to fail. In classes of usual high-school size, that is, of twenty or twenty-five pupils, or more, he believes that marks should fall within the limits suggested unless some very clear and convincing reason why they should not is apparent. Probably this reason should not only be apparent to the teacher but to the principal or other supervisory official as well. For the combined marks of all a teacher's pupils, especially during several semesters rather than for only one, he would say that they should always fall within these limits.

Although it is primarily the concern of the supervisor or administrator rather than of the teacher as to what to do if teachers do not conform to adopted distributions, it seems in place to say a few words about the matter here. A number of more or less mechanical mathematical plans for adjusting marks upward or downward, as the case may be, have been proposed. The writer does not believe in the regular use of any of these, although some of them are fairly satisfactory from the statistical standpoint. Instead every effort should be made to persuade teachers to assign marks that meet the adopted standards. If they still persist in doing otherwise and it is felt that their marks should be modified before being handed out to pupils, the official making the change may arrange them in order, in so far as this is possible, and then determine the new marks by giving to as many of those at the top as he thinks appropriate A's or whatever the highest mark is, and so on down the list. This method may also be applied by a teacher to her own marks if at the end of the semester or any other time that they are to be reported she feels that they need adjusting.

Summary.—Although some have argued that school marks should be abolished, there appear to be valid reasons why they should be retained and made as accurate as possible. As actually employed, their chief fault is their subjectivity and consequent unreliability due to the fact that different teachers base them upon many different things and have various standards in mind in giving them. This should, in so far as possible, be corrected by agreement among teachers in a department, a school, or a school system, upon the significance and basis of the marks employed, and the drawing up of sets of specifications to govern them. A number of excellent specifications of this sort have been prepared, but each group of teachers would do well to make its own. Final marks ordinarily should not depend so largely on a single examination as they do, but should give sufficient weight to daily oral and written work, laboratory work, notebooks, and so forth. Although percentile marks are most common among the wide variety employed, they are not on the whole so satisfactory as a system that employs only five or six letters or other symbols, of which two are failing. The distribution of marks according to these symbols should follow the normal curve somewhat closely for groups of several hundred, and much less so for ordinary class groups. It is better to adopt limits for the per cents of pupils who should receive each mark rather than single exact per cents.

BIBLIOGRAPHY

Blackhurst, J. H. "The Normal Curve as Related to High School and College Grading," *School and Society*, 13:447–50, April 9, 1921.

Camp, F. S. "Some 'Marks': An Administrative Problem," *School Review*, 25:697–713, December, 1917.

Courter, C. V., et al. "Uniform Marking System for the High Schools of Michigan," *Michigan Education Journal*, 3:280–81, January, 1926.

Ellis, R. S. "Converting Scores into Grades or Marks," *Standardizing Teachers' Examinations and the Distribution of Class Marks*. Bloomington, Illinois: Public School Publishing Company, 1927, Chapter V.

Jaggard, G. H. "Improving the Marking System," *Educational Administration and Supervision*, 5:25–35, January, 1919.

Johnson, F. W. "The Marking System," *The Administration and Supervision of the High School*. Boston: Ginn and Company, 1925, Chapter XV.

Kyte, G. C. "The Evolution of a Marking System from Chaos to Order," *Educational Administration and Supervision*, 6:9–16, January, 1920.

Odell, C. W. "A Selected Annotated Bibliography Dealing with Examinations and School Marks," *University of Illinois Bulletin*, Vol. 26, No. 20,

Bureau of Educational Research Bulletin No. 43. Urbana: University of Illinois, 1929. 42 p.

Odell, C. W. *Traditional Examinations and New-Type Tests.* New York: The Century Co., 1928, Chapters IV, V, and VI.

Ruch, G. M. "Examinations, Marks, and Marking Systems," *The Objective or New-Type Examination.* Chicago: Scott, Foresman and Company, 1929, Chapter XIV.

Rugg, H. O. "Teachers' Marks and Marking Systems," *Educational Administration and Supervision,* 1:117–42, February, 1915.

————. "Teachers' Marks and the Reconstruction of the Marking System," *Elementary School Journal,* 18:701–19, May, 1918.

————. "The Teachers' Use of Statistical Distributions in Giving School Marks," *A Primer of Graphics and Statistics for Teachers.* Boston: Houghton Mifflin Company, 1925, Chapter VI.

Spence, R. B. "The Improvement of College Marking Systems," *Teachers College, Columbia University, Contributions to Education,* No. 252. New York: Bureau of Publications, Teachers College, Columbia University, 1927. 89 p.

Symonds, P. M. "Marks and Marking Systems," *Measurement in Secondary Education.* New York: The Macmillan Company, 1927, Chapter XXIV.

Whitten, C. W. "Report on Standardizing Teachers' Marks," *Sixth Yearbook of the National Association of Secondary-School Principals.* Menasha, Wisconsin: George Banta Publishing Company, 1922, p. 183–202.

CHAPTER XX

NON-STANDARDIZED TESTS

The place of non-standardized as compared with standardized tests.—Although by far the larger part of this volume is devoted to standardized rather than non-standardized or teacher-made tests, this fact is not intended to imply that they are more significant from the standpoint of the teacher. On the other hand, most teachers do and should employ tests of their own construction much more frequently than commercially available ones and, therefore, should give more time and attention to them. Since, however, the writer has treated the subject at considerable length elsewhere [1] and also since so doing here would expand this volume to undesirable proportions, a relatively short treatment will be given.

In beginning this discussion it seems well to point out briefly the advantages of standardized and non-standardized tests with regard to each other. The distinguishing feature of the former, that norms have been established, is an advantage in that it renders possible the comparison of pupils with others. Norms are, however, frequently too general to be of high value and are even liable to misinterpretation and misapplication because of this fact. Also many tests may serve their purposes without such comparisons being needed.

In general standardized tests are more carefully and scientifically made than non-standardized ones. Their authors are usually better versed in methods of test construction and have a wider knowledge of subject matter than do regular classroom teachers. Therefore standardized tests conform more closely to general or best practice than ordinary examinations, and thus tend to produce uniformity. Since more time and care is devoted to their construction, they are generally more objective, reliable and valid

[1] Odell, C. W. *Traditional Examinations and New-Type Tests.* New York: The Century Co., 1928. 469 p.

than non-standardized tests, although the latter can be made equal to them if sufficient pains are taken to do so.

A third advantage of standardized tests is that their use saves time in both preparation and scoring. The objection may be raised, however, that time devoted to thoughtful preparation of test exercises is profitably spent and should not be lessened. Furthermore, the saving of time in scoring is little if any when they are compared with the best new-type tests made by teachers.

Finally, most good standard tests possess two, or occasionally more, equivalent forms, thus enabling one to test the same abilities of a group of pupils two or more times with tests of the same difficulty. This makes it possible to measure progress much more accurately than if such instruments are not available. It is difficult for teachers to prepare such duplicate forms in a manner that insures equivalence.

Probably the chief advantage of non-standard tests is that they can be much better adapted to local courses of study, to points emphasized by a teacher, to the needs of particular classes or even individuals, and so forth. In other words, they can be made to fit what has been taught rather than what someone believes should be taught, and accommodated to the exigencies of time, convenience, and so on of each teacher or school. Closely connected with this advantage, indeed one of its causes, is the fact that although the supply of standardized tests for high-school use runs well into the hundreds, there are phases of practically all subjects, and even a few whole subjects, that are as yet very inadequately covered by such tests.

From the practical standpoint the chief advantages of teacher-made tests are that they cost little or nothing except the teacher's time and that they can be prepared on relatively short notice, without planning far ahead. The first is, and probably always will be, influential enough to insure their use in a majority of cases where any test is employed. The second, however, deserves less consideration. Most standardized tests can be secured within the time for which teachers should have their work planned ahead in enough detail to know when they wish to give tests.

Another advantage ensuing from teachers making their own tests was hinted at several paragraphs above. A reasonable amount of time spent in so doing is worth while because of the more or

less enforced consideration of the aims of the course and of particular units of subject-matter, of teaching methods, of results attained, and of other similar matters which it entails.

From the consideration of these advantages on both sides and of other related points, the writer does not see how one can avoid the conclusion that no testing program for a semester, year or other large unit of work in a subject can be well balanced unless it includes both standardized and non-standardized tests. Ordinarily, if not always, the number of the latter should exceed that of the former, their respective proportions depending partly on how satisfactory is the supply of standard tests available in the subject being dealt with. Moreover use of the two varieties of tests should be so planned as to form a single unified testing program.

The relative advantages of discussion and new-type tests.—Following the conclusion that teacher-made tests should generally constitute the major part of those used the question arises as to whether they should be entirely or predominantly traditional examinations or new-type tests. Since here also there are advantages on both sides the chief ones will be presented in the next few paragraphs.

The discussion or essay examination is easier to make, especially as regards the amount of time required. The comparatively few questions of this type needed for an examination can be made in much less time than the relatively large number of new-type items. Moreover, the fact that many varieties of new-type tests need to be mimeographed or otherwise reproduced if they are to be used effectively whereas essay examinations can almost always be written on the blackboard, or even given orally, with satisfactory results, makes the latter much less trouble to employ. This very ease of use and possible rapidity of formulation, however, entails the disadvantage that teachers may give too little care and forethought to the matter with the result that a few inferior questions are dashed off at the last moment rather than a carefully considered examination constructed.

At present most teachers are more familiar with traditional than new-type tests and therefore are better qualified to make and give them. This condition is rapidly being changed, however, and will not long be a valid argument for their use. Indeed, its present validity may be questioned on the ground that the con-

struction and administration of new-type tests is not so difficult but that all teachers can learn it easily and well enough to warrant their using such tests.

It is sometimes urged that new-type tests encourage guessing more than do those of the essay type. This is probably true in practice, but the tendency in this direction can be considerably reduced if the proper directions and scoring methods are employed. Also many essay examinations encourage guessing or bluffing because of the vagueness and indefiniteness of the questions they contain.

Another argument advanced against certain forms of new-type tests is that they tend to confuse pupils by presenting false statements, incorrect answers, and similar material. There appears to be a little, but only a little, truth in this charge. If the material dealt with has been studied and learned fairly well before the test is given, the danger of confusing pupils as to what they already know is slight. With high-school pupils it is much less than with younger ones. Moreover, if the test papers are returned promptly and the correct responses emphasized most of the confused or erroneous conceptions that may have been produced will be removed and correct ones substituted.

The one most significant merit of traditional examinations is that they appear to test certain desired outcomes and mental processes better than do the other kind. Although new-type exercises may measure originality, initiative, power to organize, to interpret, to analyze and synthesize, and various other reasoning processes to some extent, they appear to do so much less thoroughly and satisfactorily than well made discussion questions. In other words, the latter possess higher validity for these purposes than do the former. They also provide opportunities such as cannot be afforded by tests to which the answers are single words or phrases, for measuring pupils' power to express their thoughts, to write well, to use correct English, and other related abilities and habits.

One of the advantages of new-type over essay examinations is implied by the expression "objective tests," frequently applied to the former. Not only can some varieties of new-type tests be made perfectly objective if sufficient care is used, but also those kinds which cannot be made so are more objective than most, indeed almost all, essay examinations. The same teacher at different

times, or different teachers at the same time, will usually agree closely if not exactly as to the scores to be given responses to a set of new-type exercises, whereas numerous published studies and still far more frequent unpublished experiences show how much disagreement there is liable to be in marking answers to discussion questions.

Partly because of their greater objectivity, and partly for another reason, new-type tests are more reliable than traditional ones. The second reason is that because of the much larger number of items which they include, they secure better samplings of pupils' abilities and knowledge than do discussion examinations of only a few questions each. This may be illustrated by the following analogy. Suppose that one wishes to determine the average intelligence of high-school freshmen in a large city which has ten high schools with five hundred freshmen in each, but that only one thousand out of the five thousand can be tested. One may choose to test all the freshmen in two schools or one hundred from each school, to state only two possibilities. If the first is done, it may easily result that because these two schools happen to draw pupils from the best sections of the city, or from the worst, from cultured homes where English is spoken or from uncultured foreign-speaking ones, or from other non-typical environments, the average secured is considerably too high or too low to represent the whole five thousand correctly. If one hundred are chosen at random from each school, all sections of the city and all degrees of intelligence are represented, hence the average secured is in all probability very close to the true average for all the freshmen.

It is claimed that new-type are more valid than essay tests. Although figures have been given to prove this contention,[2] most, if not all, of those reported are open to the criticism that satisfactory criterion measures, or standards of comparison, were not employed. Usually the criteria were composite scores on both kinds of tests, thus permitting the greater reliability of new-type tests to increase their apparent validity. Despite this fallacy, the writer believes that as general measuring instruments covering all the mental activities involved in most school subjects, new-type tests are probably more valid than the others.

Almost all published collections of pupils' opinions with re-

[2] Not all the reported studies support this argument, but most of them do.

gard to the question show that large majorities prefer new-type tests to essay examinations. The reasons appear to be their greater definiteness and objectivity, the smaller amount of writing required, that results are usually known sooner and more exactly, and that they are less strenuous and nerve-wracking. The attitude of the teacher has much influence on that of the pupils, however, and by gross misuse of either type she can render it very unpopular.

Another advantage of new-type tests is that they avoid confusion of pupils' ability to use good English and to express their thoughts in writing with their knowledge of or skill in the subject being tested. Even if they consciously attempt to do so, it is very difficult for teachers to avoid being influenced by these factors.

Somewhat similar to the point just mentioned is the fact that the scores on new-type tests made by pupils depend very slightly upon their speed of writing, as is too often the case with essay examinations. Pupils who are slow writers are at very little if any disadvantage as compared with those who write rapidly and thus earn scores which better indicate their knowledge of the subject dealt with.

A frequently mentioned and very practical advantage of new-type tests is their greater ease of scoring. Not only is less time required, but the mental work involved is not so arduous as in the case of discussion examinations. By a little forethought provision can be made that pupils' answers be placed in straight columns or otherwise so that they can be compared with a list of correct answers, similarly arranged and spaced, very quickly.

If a teacher wishes to prepare two or more similar tests of approximately equal difficulty he will probably more nearly be able to accomplish his purpose with new-type than with traditional tests. It is much more probable that the relatively large number of items in one new-type test will average about the same degree of difficulty as those in another similar one than that the same condition will hold for the few questions in two essay examinations.

To sum up the preceding discussion, it appears that new-type tests possess more advantages over essay examinations than do the latter over the former, but that a well balanced testing program will employ both. The proportions should vary according

to the type of subject matter and the desired outcomes of instruction.

General principles for the construction and administration of teacher-made tests.—Before proceeding to state principles to be applied and procedures to be employed in making examinations, it seems well to state their chief purposes. The following list of six appears inclusive enough to embrace all and yet specific enough to suggest the different functions:

The measurement of pupil ability and accomplishment
The diagnosis of pupils, especially of those doing unsatisfactory work
The measurement and improvement of teaching efficiency
The provision of opportunities for learning
The motivation of pupil study and other mental activity
The determination of standards or goals of attainment

In the following pages it will be necessary because of lack of space for the writer to state a number of points rather briefly and dogmatically without giving justification for them. If the reader desires such justification and further explanation he can find it in the source referred to at the beginning of this chapter.[8] As was suggested at the beginning of Chapter III, a number of the criteria for standardized tests apply to non-standardized tests as well; therefore those already discussed in that chapter will receive little attention here.

A good examination should be highly objective so that there will be no doubt or as little as possible as to what the correct answers are.

Another desirable quality of non-standardized as well as standardized tests is reliability. The teacher ordinarily cannot try out two forms of a test before using either, but should endeavor to secure high reliability by length, form, selection of items, conditions of administration, and other similar factors affecting it.

It is, of course, of first importance that a non-standardized test as well as any other be valid, that it measure the thing it is intended to measure. The chief means to be employed by the teacher in securing validity is careful selection of test items and their inclusion in exercises so formulated that pupils will have no doubt

[8] Odell. *Op. cit.*

as to their meaning. If teachers prepare examination exercises several days in advance of the time they are to be used, then lay them aside and examine them again, it is probably that greater validity will result. Suitable questions and exercises should be recorded from time to time as they occur. They may be parts of previous tests that have proven satisfactory, or of examinations made by others, questions handed in by pupils, exercises that occur to the teacher in connection with daily work, and so forth.

A good examination should be easy for the teacher to give and score and likewise for the pupils to take. Also it should be economical of the time of both, not requiring that pupils devote an undue amount of time to it nor that the teacher be heavily burdened by its preparation, administration, and scoring.

The exercises contained in a test should be such, both in form and content, as to arouse interest on the part of the pupils and stimulate them to put forth approximately their best efforts.

Catch questions should rarely be used. An occasional one is probably justified in high school, although some teachers would absolutely abolish them.

Some examinations should be either long enough or difficult enough that no or practically no pupils are able to complete them with perfect scores and likewise short or easy enough that all pupils are able to answer one or more exercises correctly. Probably all examinations should conform to the last requirement but not to the first, since in the testing of minimum essentials and other more or less similar content examinations should sometimes be given upon which it is expected that a number of pupils will make perfect or near-perfect scores.

Pupils should not be permitted choice of questions. If, however, for any reason a teacher is determined to allow such choice, it should not be such that pupils can omit all exercises of a particular variety. For example, in a foreign language examination they should not be allowed to omit all translation, but perhaps to omit one out of several passages given, or in a history examination not to omit all dates, but perhaps to give eight out of ten or some other similar proportion.

There should be no exemptions from examinations. Many schools allow them, but practically all advantages gained appear to be procurable in other ways, thus avoiding the definite disadvan-

tages that result. The chief of these disadvantages are that such a system tends to make examinations a penalty and deprives pupils of benefits that may be gained from them.

There should be wide variety in examinations. Some should be written, some oral; some should be traditional, some new-type; some should be long, some short; some should depend largely upon speed, others should not; some should test general principles, others detailed facts; some should measure knowledge of theory, others ability to apply; some should be announced in advance, some should not; on some tests pupils should be allowed to use their textbooks and other helps, on others, probably most, they should not. In other words, the teacher should attempt to measure as many phases as possible of the knowledge and abilities supposed to have been acquired by pupils.

Examinations should be constructed so as to discourage bluffing or guessing, and to encourage regular study and true review as distinguished from cramming. Likewise they should, of course, be such that cheating is made as difficult as possible. Both form and content should be chosen with these aims in mind.

Examination questions and exercises should be clearly and definitely stated and accompanied by directions that are likewise clear and definite and as brief as is consistent with these qualities. If the types of exercises used are not familiar to pupils, examples or practice exercises should precede the test itself.

The times at which tests are given should be determined by the natural divisions of the subject, the progress of the class, and the remainder of the teacher's instructional program rather than by the periods at which reports are issued or other extraneous circumstances.

No one test given high-school pupils should be longer than an hour and a half, and those of this length should be rare, probably not oftener than once a semester. There should usually be two or three more per semester lasting the whole recitation period and many shorter ones. On most, perhaps all, tests of more than fifteen or twenty minutes in length, pupils should be given warning a few minutes before time is up. The best time to give this is perhaps when there is from 10 to 20 per cent of the total time yet remaining. On discussion examinations containing questions that call for general planning and organizing pupils should be encouraged

not to begin writing immediately, but to take sufficient time to plan their answers first.

If an examination consists of many questions or elements, it is best that a copy be placed in the hands of each pupil. If it is written on the board instead, the teacher should read it so as to avoid any danger of pupils' misunderstanding her writing.

If pupils have not exactly followed directions, they should ordinarily be given credit for their evident intentions and the substance of their answers. For example, if they are told to underline certain words out of a list, and instead check them, they should be allowed credit if the words checked are the correct ones.

In such tests as single-answer, multiple-answer, and so on, which offer answers from which the correct ones are to be selected and indicated by the pupils, care should be taken that the arrangement of the correct answers among the others is random or irregular. Also in such tests the incorrect answers should not be too evidently wrong, but should be such as tend to mislead pupils who know nothing or little about the subject dealt with. How nearly correct they are largely determines the difficulty of a test.

Scoring pupils' responses.—One of the most important principles relating to this is that ordinarily pupils' papers should be scored and returned to them as promptly as possible, while the test is still relatively fresh in their memory and their interest still high. Before being returned the papers should be carefully marked, all errors being indicated. This does not mean that they should be corrected; indeed, it is usually better not to do this, but to require the pupils to do so individually or else in class discussion.

Returned test papers should be so scored that pupils know definitely what they have earned upon them. Furthermore, they should be given information concerning the scores of the rest of the class such that each can determine approximately how he stands with regard to the rest of the group. One easy but reasonably satisfactory way of accomplishing this is to tell the pupils the highest, middle, and lowest scores made.

In the case of traditional examinations containing comparatively few questions, it is probably best to weight the questions on the basis of their estimated importance. For new-type tests con-

taining fairly large numbers of items, weighting is ordinarily not worth the extra trouble it entails, and it is best to count one or occasionally some other number of points upon each response called for.

It is usually more than worth the time it takes for a teacher to prepare a list of correct answers to a test before she begins to mark the pupils' papers. Indeed, if these are prepared at the same time that the examination is constructed, it will often result in improving the examination and lessening her labor later by causing her to eliminate or revise questions that would be hard to score. As decisions are made concerning doubtful or partial answers she should record them so as to be sure other pupils who respond similarly are given the same scores thereon.

In scoring discussion examination papers, it is usually well to do all the answers to one question at a time, thus keeping one's attention fixed thereon, and so probably maintaining a more uniform standard or rating.

In handling questions to which the answers are rather general, not being based on a specific number of points, it is frequently desirable to use the so-called sorting method. This involves the sorting or classifying of the answers into a number of groups or piles each containing those of approximately the same degree of merit. Frequently five piles is a satisfactory number.

In scoring papers in other subjects than English a teacher should follow certain rules concerning what should be done about errors in capitalization, punctuation, grammar, spelling, poor handwriting, lack of neatness, and so forth. Indeed, not merely the individual teacher, but the department, or better yet the school, should follow a uniform procedure in this respect. Ordinarily it is best not to give lower marks to pupils' responses in the various school subjects because of poor English, and so forth, but to mark the errors in English and deal with them separately.

In scoring pupils' responses to relatively important examinations, particularly those at the ends of semesters and years, it is often desirable to use the committee system of marking, provided there are two or more teachers of the same subject in the school. According to this system one teacher marks all of the answers to one question, another teacher all of those to another question, and

so on. This secures greater reliability of marking by averaging the idiosyncrasies of individual teachers, but has the disadvantage that unless each teacher makes another inspection of her pupils' papers she does not know just what they have done as well as if she graded the entire papers.

In many cases it is desirable to have pupils score their own papers or exchange them and score one another's. By having them do this under supervision much useful practice and drill may be afforded them and the burden of dealing with their written work considerably lessened.

For scoring almost all varieties of new-type tests, answer keys containing the correct responses in straight columns or otherwise arranged so as to be most convenient should be prepared. By their use the scoring of such tests can be done very rapidly.

Traditional, essay, or discussion examinations.—Since practically every teacher is much more familiar with this variety of examination than with new-type tests, it will be given very limited attention here. This is not meant to imply that it is of slight importance, since, as has already been stated, the writer believes that a fairly large proportion of all examinations should be of this type. Essay examinations should rarely be employed to test knowledge of isolated facts and details and other outcomes of instruction that can be satisfactorily dealt with by the use of new-type tests, since the latter have greater objectivity, reliability, and other advantages, but for the testing of numerous mental activities, such as are mentioned in the next few pages, they have values not possessed by new-type tests.

Several writers have suggested the different types of mental activity that may well be covered by test questions. The list formulated by Monroe and Carter,[4] although not complete, is the best with which the writer is familiar. The twenty types that this contains and also two questions or exercises illustrating each are presented below.[5]

[4] Monroe, W. S. and Carter, R. E. "The Use of Different Types of Thought Questions in Secondary Schools and Their Relative Difficulty for Students," *University of Illinois Bulletin*, Vol. 20, No. 34, Bureau of Educational Research Bulletin No. 14. Urbana: University of Illinois, 1923. 26 p.

[5] These questions and exercises are quoted from: Odell, C. W. *Traditional Examinations and New-Type Tests.* New York: The Century Co., 1928, p. 207–10.

1. Selective recall—basis given.
   Name the presidents of the United States who had been in military life before they were elected.
   What do New Zealand and Australia sell in Europe that may interfere with our market?
2. Evaluating recall—basis given.
   Which do you consider the three most important American inventions in the nineteenth century from the standpoint of expansion and growth of transportation?
   Name the three statesmen who have had the greatest influence on economic legislation in the United States.
3. Comparison of two things—on a single designated basis.
   Compare Eliot and Thackeray as to ability in character delineation.
   Compare the armies of the North and South in the Civil War as to leadership.
4. Comparison of two things—in general.
   Compare the early settlers of the Massachusetts Colony with those of the Virginia Colony.
   Contrast the life of Silas Marner in Raveloe with his life in Lantern Yard.
5. Decision—for or against.
   Whom do you admire more, Washington or Lincoln? Why?
   In which in your opinion can you do better, oral or written examinations? Why?
6. Causes or effects.
   Why has the Senate become a much more powerful body than the House of Representatives?
   What caused Silas Marner to change from what he was in Lantern Yard to what he was in Raveloe?
7. Explanation of the use or exact meaning of some phrase or statement in a passage.
   Explain the meaning of the expression "Sinais climb" in the line: "We Sinais climb and know it not."
   Explain the meaning of the word "original" in the statement: "The Supreme Court has original jurisdiction only in cases wherein a state or diplomatic representative is a party."
8. Summary of some unit of the text or of some article read.
   Summarize in about one hundred words the advantages of the hot-air furnace.
   Summarize in not more than one page what is to be found in the text concerning reconstruction in the South after the Civil War.
9. Analysis.
   What characteristics of Silas Marner make you understand why Raveloe people were suspicious of him?
   Mention several qualities of leadership.
10. Statement of relationships.
    Why is a knowledge of botany helpful in studying agriculture?
    Tell the relation of exercise to good health.

11. Illustrations or examples (your own) of principles in science, construction in language, etc.
    Give an original sentence in Latin illustrating the use of the infinitive in indirect discourse.
    Show how some one phenomenon with which we are familiar in everyday life illustrates the fact that heat commonly causes expansion.
12. Classification. (Usually the converse of No. 11.)
    What is the construction of the word "me" in the sentence "The boy gave me his book"?
    To what group of plants do the mosses and liverworts belong?
13. Application of rules or principles in new situations.
    In what countries other than Brazil would you expect to find rubber plantations?
    What chemical properties would you expect phosphorus to possess, knowing that its position in the periodic table is below that of nitrogen?
14. Discussion.
    Discuss the Monroe Doctrine.
    Discuss early American literature as best you can in about two hundred and fifty words.
15. Statement of aim—author's purpose in his selection or organization of material.
    What was the purpose of the author in having Athelstane return to life after he was apparently dead?
    Why do you suppose the author of this history relates the story of Barbara Frietchie when her act had no bearing on the outcome of the war or even of any particular battle?
16. Criticism—as to the adequacy, correctness, or relevancy of a printed statement, or a classmate's answer to a question on the lesson.
    Do you believe that the following statement is true? Give your reasons. "The South would have won the Civil War if it had possessed an adequate navy."
    Criticize "Macbeth was wholly indifferent to the superstitions of his time."
17. Outline.
    Outline in not more than one page the chief events of the French and Indian Wars.
    Outline the Constitution of the United States, including the amendments passed very shortly after the adoption of the Constitution.
18. Reorganization of facts.
    Trace the life of a glacier from its beginning, showing the steps in its origin, its development, etc., down to its destruction.
    Select the incidents that characterize Portia in the *Merchant of Venice*.
19. Formulation of new questions—problems and questions raised.
    In addition to what is stated in your text, what other questions can you think of which need to be answered to explain why some portions of the earth's surface are higher than others?
    If you were asked to state how much you could trust the viewpoint of

a particular historian about whom you know little or nothing, what questions would you want to have answered concerning him?

20. New methods of procedure.

How might the plot of *Julius Cæsar* be changed to make it a comedy rather than a tragedy?

How would you prove or disprove the statement that oil is a more efficient heating fuel than soft coal?

Single-answer tests.—The simplest and probably the most familiar variety of new-type tests is the single-answer or direct-recall exercise. It consists of a direct question or its equivalent to which the answer is a single word, number, or expression of some sort. Such exercises have long been employed by teachers, but with the recent emphasis on new-type tests, they have received much more attention in the past few years than formerly. Because of the fact that such exercises have been long used, are relatively easy to construct and also for the pupils to understand, this type is perhaps the best to be used at first by teachers who are not familiar with new-type tests. Furthermore, it is among the three or four most generally useful varieties of such tests.

The most convenient form in which to arrange such a test is to have the questions composing it preceded by blank lines forming a straight column at the left or else followed by lines forming a column at the right on which pupils are to record their answers. This form is illustrated by the example given below.[6]

Directions. The correct answer to each question below is a single word. If you know, or think you know, the answer to a question, write it upon the blank line in front of that question. Do not write more than one word upon any line.

_____1. What term is applied to all money and business papers that pass as money?

_____2. What is an itemized statement of amount and value of goods on hand called?

_____3. What must be deducted from the total receipts to find the net profit?

_____4. What kind of a draft is payable on presentation?

_____5. What name is applied to cashing a note before it is due?

[6] Because of lack of space no attempt will be made to illustrate or even mention all of the minor varieties of new-type tests. Instead short examples will be given of one or two varieties of each of the chief types, those being chosen that are most appropriate for general use and serve to illustrate the type represented best, and other possibilities mentioned.

One variety of single-answer tests is that in which a list of terms, each to be followed by a corresponding term of some sort, is given. Thus, for example, abbreviations of musical terms or chemical symbols may be given to be expanded, or the complete expressions given with abbreviations or symbols called for. In cooking a list of foods may be given with the time that each should be cooked or the per cent of protein in each. Other possibilities are to give a list of foreign words and ask for their English equivalents, or vice versa, to call for the dates of historical events or the outstanding occurrences in the lives of historical personages. Another kind gives definitions or descriptions and calls for the things defined or described. For example, in literature there may be a series of short sketches, perhaps two or three sentences each, of a number of writers and the pupils asked to identify the persons referred to. Still another variety calls for several similar responses, each a single expression and more or less independent of the others, to each item or question. Still different is the type that asks for two or more responses of different sorts. Thus it may contain quotations for which pupils are to supply the authors and works, or, if from plays, the characters who spoke them.

Multiple-answer tests.—For all-around use the writer is inclined to favor the multiple-answer, also called multiple-choice, multiple-response, best-answer, and so forth, test over the other new types. Its essential feature is that it presents several possible answers to pupils from among which they are to select the one or perhaps more correct ones. The number of answers varies, but four or five are ordinarily best. Such tests may be adapted to any school subject and almost any phase thereof. Furthermore, if the suggested answers are carefully selected, they obviate the possibility of doubt as to whether pupils' answers are correct or not that may exist in the case of single-answer and some other varieties. Their difficulty can also be rather easily regulated by the fineness of discrimination required to select the correct answers. The objection has been raised to them that they tend to confuse pupils by putting incorrect answers before them, but it appears that if they deal with content previously studied, there is very little of this effect, at least for individuals as mature as high-school pupils. It is undoubtedly true that they require somewhat less initiative than the single-answer and some other types, but

if they are not overused their advantages more than balance this objection.

There has been considerable discussion concerning the method of scoring multiple-answer tests, but the ordinary teacher employing them need not pay much attention to it. If they contain four or more suggested answers the method of merely counting the number of correct answers as a pupil's score is satisfactory enough for classroom use.[7]

It is desirable in multiple-answer as in single-answer and other forms of tests to have the pupils' responses appear in a straight column. Usually the best means of accomplishing this is to number or letter the suggested answers and have the proper numbers or letters copied on short blanks at the left or perhaps at the right of the test sheet. A portion of such a test is illustrated below.

Directions. Each of the questions below is followed by five suggested answers of which one is right. If you think you know which one is right, place the letter before it on the short blank line in front of that question.

_____1. Who wrote *The House of Seven Gables*? A. Irving, B. Thackeray, C. Poe, D. Hawthorne, E. Stevenson.

_____2. Who was the author of *The Passing of Arthur*? A. Tennyson, B. Browning, C. Goldsmith, D. Kipling, E. Macaulay.

_____3. In which country does most of the action of *Ivanhoe* occur? A. Scotland, B. Ireland, C. France, D. Wales, E. England.

_____4. What is the name of the merchant in *The Merchant of Venice*? A. Orlando, B. Bassanio, C. Prospero, D. Antonio, E. Oliver.

_____5. In which selection is Gabriel an important character? A. *Hiawatha*, B. *Treasure Island*, C. *Evangeline*, D. *Lorna Doone*, E. *The Scarlet Letter*.

There are a great many forms of multiple-answer tests, some of which have no virtue except that their use provides variety. Others, however, are especially adapted to certain content or cer-

[7] The formula Score $= R - \frac{W}{N-1}$ allows for guessing and thus yields more accurate scores in a certain sense than merely counting the number right. The validity and reliability of scores computed on this basis are so slightly greater than those from merely the number right that it does not seem worth while to go to the trouble of employing it ordinarily. In this formula "R" equals the number of right answers, "W" the number of wrong answers, and "N" the number of suggested answers to each element.

tain purposes of testing, and a number of these will be mentioned. In one the difference consists in the fact that there is one incorrect answer in the group, the majority being correct. An example of this is:

———————1. Lived in eighteenth century. A. Louis XVI, B. William Pitt, C. Charles II, D. Frederick the Great, E. Maria Theresa.

For this, of course, *C* is the correct response since Charles II was the only one of the persons named who did not live in the eighteenth century. Another variation is that there may be more than one correct answer among those given. For example, in the last there might have been two persons included who did not live in the eighteenth century or, if the directions had called for indicating persons who did live in that century rather than those who did not, there might have been any number of the five who did. Sometimes the suggested answers are not single words or short expressions, but definitions, explanations, or reasons. Thus in geometry, for example, such terms as "plane," "hypotenuse," "isosceles triangle," and so forth, may be given with a number of definitions for each of which only one is correct. In foreign language there may be a number of translations of a short passage, or a number of grammatical explanations of a construction, of which one is to be indicated. Sometimes there are merely groups of expressions without any preceding question or statement which are to be marked on the basis of indicating one or more not associated with the rest, one that is the cause whereas the others are results or vice versa, or in some other more or less similar fashion. For example, in agriculture one such exercise might contain the terms "Arabian, Galloway, Clydesdale, Belgian, Percheron." In this, of course, "Galloway" is to be marked, since the other four refer to horses, but it does not. Also in history such groups may be used as: "Fall of Constantinople, discovery of America, invention of the compass, revival of learning." In this "discovery of America" is the result and the other three causes. Sometimes instead of having a separate set of suggested answers for each question there is a single set for a number of questions or other elements. Thus, for example, a series of historical events or characters may be named to be classed in some one of several periods of time, or as belonging to some one of several nations.

What is sometimes called the compound multiple-answer test differs from those previously mentioned in that it calls for two or more responses to each exercise on different bases. Thus from groups of historical characters or events the earliest and the latest in point of time may be called for, in foreign language work the subject and object of each sentence, and so forth. A variation is that in which the same exercise contains two sets of several answers each and requires that the correct one from each set be indicated. For example, in such an exercise as:

"The (cell, protoplasm, molecule, atom), the primal material of life, is a (jelly-like, relatively hard, liquid) substance," the term "protoplasm" from the first group, and "jelly-like" from the second, are the only two that will make a correct statement. In foreign language there may be groups of several English and several foreign words with one of each group meaning the same as one of the other. For example, in French we might have such a group as: "sick, well, lazy—loin, malade, lent," in which "sick" and "malade" mean the same.

Alternative tests.—Although alternative tests, that is, tests with only two suggested responses, are really a variety of multiple-answer tests, they have received so much attention as to seem to merit separate treatment. They are relatively easy to construct, can be answered rapidly, and call for a certain kind of critical ability that is worth testing.

There has been even more contention as to how such exercises should be scored than in the case of multiple-answer tests with several suggested answers. It will be seen that the formula previously given, $\text{Score} = R - \frac{W}{N-1}$, reduces to $\text{Score} = R - W$, when $N = 2$, as is the case with alternative tests. In other words, the score is simply the number right minus the number wrong, paying no attention to those omitted. There is not space here to enter into the arguments for and against the use of this formula. Suffice it to say that the validity of scores determined by it appears to be slightly greater than that of merely the number right, and also that the knowledge that such a method of scoring is used tends to exercise some desirable restraint upon guessing on the part of pupils. Therefore the writer recommends that it be used.

The most common variety of alternative tests is the true-false

exercise consisting of a series of statements, some of which are true and some of which are false, usually in about equal numbers. An example of such a test is given below.

Directions. Below are a number of statements of which about half are true and the other half false. In the case of each statement that you think is true, place a plus (+) mark on the blank line in front of it, and in the case of each that you think is false, place a minus (−) sign. To show just how this is to be done, the first two sentences have been marked. The first one is true, so it has a plus sign in front of it, and the second is false, so it has a minus sign before it. If you do not think you know whether a statement is true or false, do not guess, but omit it and go on to the next one.

+ 1. Steam may be defined as water vapor at the boiling point.
− 2. A prism separates a shaft of white light into eight distinct colors.
____ 3. Most of the evergreens have soft wood.
____ 4. When air is heated it expands.
____ 5. The earth is larger than either Mars or Venus.
____ 6. Granite is a relatively durable rock.

Although the true-false type of test is more commonly used, the writer believes that, if an alternative test is to be employed, yes-no questions are preferable. Although there is probably little confusing effect in seeing false statements about matters that have already been studied, yet there seems to be even less if yes-no questions are used. Thus the statements in the example just given might be reworded as follows:

____ 1. May steam be defined as water at the boiling point?
____ 2. Does a prism separate a shaft of white light into eight distinct colors?
____ 3. Do most of the evergreens have soft wood?

and so on.

A modification of the alternative test that in many cases is an improvement provides a third possibility. Usually the third possible answer is "doubtful," "sometimes," or something equivalent thereto. Such a test, of course, contains some statements that are always true, some that are always false, and some of which the truth is not known or that are sometimes one and sometimes

the other. A variety of this may be employed in vocabulary work either in foreign language or English. This consists of presenting pairs of words to be marked according to whether they mean the same, the opposite, or neither the same nor the opposite.

Completion tests.—A completion test consists of statements, sometimes sentences and sometimes longer passages, with certain words left out for pupils to supply. Sometimes no help is given, sometimes the words to be supplied are to be taken from a list. In the first form completion tests are very similar to single-answer tests, the chief difference merely being in the wording of the statements and the positions of the omitted words. If no list of suggested answers is given, pupils are almost certain to supply some concerning the correctness of which it is difficult to decide. It is, therefore, recommended that ordinarily lists of answers be provided. Furthermore, there should ordinarily not be a large number of omitted words, especially in succession to one another. One or at most two to the sentence will usually be enough. Generally these should be the most important words.

The ordinary method of scoring completion tests is to count one point on each blank rather than on each sentence or statement to be completed.

Two examples are given below. The first is an ordinary completion test, in paragraph form, without suggested answers, and the second one with a list of selected answers.

Directions. Each of the blanks in the paragraph below represents the omission of one word. If you think you know the word that should be there, write it on the blank. Do not in any case write more than a single word on one blank.

Our federal government has three branches, the _____, the _____, and the _____. The President is at the head of one branch, Congress at that of another, and the _____ _____ at that of the third. The President is assisted by his _____, in which there are _____ members. Congress consists of _____ senators from each State and representatives whose number is determined by the _____ of the various States.

Directions. Below are a number of statements with the names of certain characters omitted. The names of these characters and also of several others are given in a list at the right. If you know it, write the correct name on each blank. If you do not think you know which is correct, do not guess. Do not use any name more than once.

1. When Washington was elected President, _____ became Vice President.	Adams
2. The Secretary of the Treasury under Washington was _____	Burr
3. Our treaty with England, made during Washington's administration, was negotiated by _____.	Hamilton
4. _____ succeeded Adams as President.	Jay
5. Washington's first Secretary of War was _____.	Jefferson
	King
	Knox
	Monroe

In the sample above there are more answers than pupils are to use so as to prevent their determining those to any of the exercises by elimination if they know most of them. Sometimes the same number of answers as blanks is given, but the writer believes this undesirable. Another variation is to permit each answer to be used as often as needed, some perhaps being used once, some twice, some even more. If this is done, the number of answers may be decidedly less than the number of blanks.

Matching tests.—A matching test presents two, or occasionally more, sets or lists of items or expressions to be matched with one another. For example, there may be one list of dates and another of events, one of Latin words and another of their English equivalents, one of terms defined and another of definitions, one of authors and another of their works, and so on. Such tests fit almost all subjects, although not all portions thereof; and where they are appropriate, constitute one of the few best types to employ.

In constructing a matching test the two or more lists may include the same or different numbers of items. The first is subject to the same objection as in the case of completion tests, that if pupils know most of them they may determine others by elimination. It is, therefore, recommended that one list contain several more than the other. The shorter list should rarely if ever contain more than a dozen items, the longer one perhaps fifteen to eighteen. If there are more, too much time may be wasted by pupils in merely looking up and down the lists. Furthermore, it is desirable that the items in one list, at least, be arranged in some regular order, such as alphabetical or chronological, that is random in regard to the other list, and yet enables pupils to find the various items readily.

Two examples of matching tests are given below. The first is an ordinary one in which there are two columns, and the second

what is sometimes called a compound matching test composed of three columns.

Directions. Below you see at the left a list of Latin words and at the right one of English words. One of the English words is the correct translation of each Latin word. Indicate which it is by placing the letter immediately before it on the short blank line in front of the Latin word. Do not use all of the English words and do not use any one more than once.

____ 1. accedo	a. approach
____ 2. contendo	b. assemble
____ 3. convenio	c. carry
____ 4. discedo	d. depart
____ 5. invenio	e. find
	f. hasten
	g. live
	h. think

Directions. Below are three lists, one of historical characters, another of countries, and another of events. If you know the country from which each character came and the event in which he was prominent, place the letter found in front of the country and the figure in front of the event on the short blank line in front of his name. There are more countries and events named than there are characters, so do not attempt to use them all, also do not use any one more than once. What you are to do is illustrated by the first named. James I came from Scotland and participated in the accession of the Stuarts, therefore *H* and *1* are placed on the short line before his name.

H 1 I. James I	A. Austria	1. Accession of Stuarts
____ II. Louis XIV	B. England	2. Expulsion of Stuarts
____ III. William of Orange	C. France	3. French Revolution
	D. Italy	4. Holy Alliance
____ IV. Metternich	E. Netherlands	5. Hungarian Revolution
____ V. Frederick the Great	F. Prussia	6. Rise of Russia
	G. Russia	7. Thirty Years' War
____ VI. Gustavus Adolphus	H. Scotland	8. War of Austrian Succession
	I. Sweden	9. War of Spanish Succession

Incorrect statements tests.—This type of test is somewhat similar to the true-false, consisting of statements of which a number have been made incorrect in such a way that by the change, insertion, or removal of a word or sometimes more, they can be made correct. Such tests are relatively difficult to construct so as to insure that there are no other changes that may be made that will cause the false statements to be true or doubtful, thus causing

difficulty in scoring. They have, however, a certain value in testing critical ability that seems to justify their occasional use. They should probably not be included among the three or four varieties of new-type tests to be employed most frequently. In constructing them any desired proportion of the statements included may be made wrong. Sometimes even all of them are so, but usually some are left correct. Sometimes it is only required that the incorrect words be indicated, but it is probably better to require that the correct ones be supplied also. An example follows.

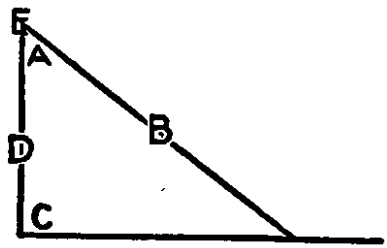
Directions. Some of the statements below are true; others are false because of some one incorrect word that each contains. If you think a statement is true, write the word "true" on the blank line in front of it. If you think it is not true, draw a line through the one word that makes it false and write the correct word on the blank line in front of it.

_____ 1. Scissors are four inches or less in length, whereas shears are longer.
_____ 2. Gingham is made of cotton.
_____ 3. The warp is the piece of cloth that runs across the piece.
_____ 4. Printing usually results in a less lasting color than other methods of dyeing.
_____ 5. A French seam is one so made that the raw edge of the cloth is exposed.

Identification tests.—Identification tests are essentially matching tests in which instead of one series of items there is a figure of some sort to be matched with a series of terms. For example, a map may be numbered or lettered and a series of names of cities, countries, products, or something else be given to be matched with the indicated locations on the map, or a figure of an animal, a plant, or something else may be so marked and a list of terms given to be matched with the marks. In such tests, as in matching tests, it is preferable not to have just the same number of symbols on the figure and of terms to be matched with them, but to have several more in one list than in the other. An example of this type of test as it may be employed in geometry is given below.

Directions. You see below a triangle with several of its parts lettered, also a list of geometrical terms. This list includes all of the lettered parts of the triangle and also some others. Place each letter found on the triangle on the short blank line in front of the proper expression.

_____ 1. acute angle
_____ 2. exterior angle
_____ 3. hypotenuse
_____ 4. leg
_____ 5. obtuse angle
_____ 6. right angle
_____ 7. supplementary angle
_____ 8. vertex



Sometimes the process is reversed and the numbers or letters are given with the terms and are to be placed upon the figure. This is usually undesirable, however, since pupils are likely to be so inaccurate in placing them that it is doubtful whether or not their knowledge is correct.

Rearrangement tests.—Rearrangement or continuity tests present series of items in confused or random order and require pupils to rearrange them in some designated order. For example, historical characters or events may be given to be arranged in chronological order, physical substances to be arranged in order of specific gravity, liquids in order of boiling or freezing points, foods in order of time required to digest or per cent of fat, and so on. There are many situations in which such tests are not appropriate, but where they are the writer believes that they may well be used frequently. In constructing such tests the number of items in a single list should not be very great. Probably five or six is a desirable number, with ten as a maximum. Scoring is slightly involved, since undoubtedly pupils should receive more credit for having items almost in the correct places than nowhere nearly there, but yet not as much as if they are placed entirely correctly. Several methods of scoring have been proposed, no one of which is free from objections. One is as follows. Let the total number of possible points, that is, the number given for an entirely correct answer on any particular exercise, be equal to the greatest possible sum of differences, that is, the sum of the differences between the correct order and the most incorrect order

possible. These sums for numbers of items from four to ten are as shown below.[8]

Number of items	*Sum of difference*
4	8
5	12
6	18
7	24
8	32
9	40
10	50

The sum of the differences between the order actually given by a pupil and the correct order should be found and subtracted from the greatest possible sum of the differences, and the remainder taken as his score. Suppose, for example, that an exercise gives the names: Clay, Garfield, Jefferson, Lincoln, and McKinley. The proper order is, of course, Jefferson, Clay, Lincoln, Garfield, McKinley. Suppose, however, a particular pupil arranges them thus: Jefferson, Lincoln, Garfield, Clay, McKinley. He has Jefferson first where he should be so that there is no difference; Lincoln is second but should be third, a difference of one; Garfield is third but should be fourth, also a difference of one; Clay is fourth instead of second, a difference of two; McKinley, like Jefferson, is correct. The sum of the pupil's differences is, therefore, $1 + 1 + 2$ or 4. The greatest possible sum of differences for five items is shown by the table to be 12. Subtracting 4 from this, the pupil's score is 8.

An example of this type of test as it might be employed in cooking is given below.

Directions. Below are a number of groups each containing the names of six articles of food. In front of each group is a statement which directs that the terms composing that group be numbered in a certain order. For example, in Exercise 1 they are to be numbered in order of protein content beginning with the one that has the greatest. Therefore in this exercise you should place a figure *1* in front of the article of food that has the greatest protein content, a figure *2* in front of the one that ranks second, and so on down until you place a figure *6* in front of the one that has the least protein content. When you have completed Number 1 go ahead and do each

[8] If the number of items is even, the sum of the differences is equal to one-half the square of the number; if it is odd, the sum equals one-half of one less than the square.

of the others, numbering them according to the basis indicated in the statement.

1. Number in order of protein content, beginning with the one that has the greatest.	white bread eggs round steak spinach butter potatoes
2. Number in order of per cent of carbohydrate, beginning with the largest.	sugar potatoes oatmeal lima beans rice canned salmon
3. Number in order of proportion of water contained, beginning with that having the greatest.	tomatoes cabbage lettuce watermelon cucumbers radishes
4. Number the stages in cooking sugar in order of temperature, beginning with the lowest.	small thread caramel crack blow soft ball pearl
5. Number in order of per cent of mineral matter, beginning with the greatest.	dried beans lean beef cheese carrots milk spinach

Summary.—This chapter begins by discussing the place of non-standardized as compared with standardized tests and attempts to show that both have a legitimate place in high-school measurement, and that neither should displace the other. Certain advantages accrue from each which the other cannot supply. The same is also true of traditional or discussion examinations and new-type tests. Following this is a statement of the purposes to be served by examinations and a number of principles and suggestions for constructing, giving, and scoring them. Traditional examinations and eight varieties of new-type tests: single-answer,

multiple-answer, alternative, completion, matching, incorrect statements, identification, and rearrangement are each discussed briefly and illustrated. Of the eight varieties of new-type tests just named, the single-answer, multiple-answer, alternative, completion, and matching should probably receive the widest use; the identification and rearrangement are also good but less generally appropriate; incorrect statements are probably least valuable.

BIBLIOGRAPHY

Buckingham, B. R. "New-Type Examinations," *Research for Teachers.* New York: Silver, Burdett and Company, 1926, Chapter VI.

Cook, C. G. *New Type Questions in Chemistry.* New York: Globe Book Company, 1927. 106 p.

Ellis, R. S. *Standardizing Teachers' Examinations and the Distribution of Class Marks.* Bloomington, Illinois: Public School Publishing Company, 1927. 170 p.

Hopkins, L. T. (Edited by). *The Construction and Use of Objective Examinations.* Boulder: University of Colorado, 1926. 119 p.

Miller, G. F. *Objective Tests in High School Subjects.* Norman, Oklahoma: George Frederick Miller, 1926. 168 p.

Odell, C. W. "A Selected Annotated Bibliography Dealing with Examinations and School Marks," *University of Illinois Bulletin,* Vol. 26, No. 20, Bureau of Educational Research Bulletin No. 43. Urbana: University of Illinois, 1929. 42 p.

———. *Traditional Examinations and New-Type Tests.* New York: The Century Co., 1928. 469 p.

Orleans, J. S. and Sealy, G. A. *Objective Tests.* Yonkers, New York: World Book Company, 1928. 373 p.

Paterson, D. G. *Preparation and Use of New-Type Examinations.* Yonkers, New York: World Book Company, 1925. 87 p.

Ruch, G. M. *The Improvement of the Written Examination.* Chicago: Scott, Foresman and Company, 1924. 193 p.

———. *The Objective or New-Type Examination.* Chicago: Scott, Foresman and Company, 1929. 478 p.

Ruch, G. M. and Rice, G. A. *Specimen Objective Examinations.* Chicago: Scott, Foresman and Company, 1929.

Ruch, G. M. and Stoddard, G. D. "Informal Objective Examination Methods," *Tests and Measurements in High School Instruction.* Yonkers, New York: World Book Company, 1927, Part III.

Ruch, G. M., et al. *Objective Examination Methods in the Social Studies.* Chicago: Scott, Foresman and Company, 1926. 116 p.

Russell, Charles. *Classroom Tests.* Boston: Ginn and Company, 1926. 346 p.

Smith, Gale. *How to Construct and Use Non-Standardized Objective Tests.* Fowler, Indiana: Benton Review Shop, 1929. 168 p.

Smith, Gale. *Twentieth Century Practice Exercises and Objective Tests in Physics.* Fowler, Indiana: Benton Review Shop, 1929. 113 p.

Smith, H. L. and Wright, W. W. "New-Type Examinations," *Tests and Measurements.* New York: Silver, Burdett and Company, 1928, Chapter XXI.

Symonds, P. M. *Measurement in Secondary Education.* New York: The Macmillan Company, 1927, Chapters III and XXV.

CHAPTER XXI

CLASSIFICATION AND PROMOTION

Introduction.—Although the topics of classification, including certain types of provisions for differences in ability, and of promotion are frequently treated separately, they are so closely connected and overlap in so many points that it seems better not to attempt to divide them. It is true that formerly, and unfortunately even yet in many school systems and in the thinking of many teachers, pupils were either promoted or failed at the end of a semester or year with no thought that anything else was desirable or even possible. On the other hand, however, any system of classifying or grouping pupils to provide for individual needs and capacities must take into account promotion and failure. The whole problem, as several writers have already pointed out, should be considered one of pupil placement, that is, of determining where each pupil can best be placed so as to derive the most profitable results from his school work and then putting him there. Although this does not necessarily involve the use of standardized, or even of any, tests, they actually play such a large part in the placement of pupils that a discussion of the whole subject seems in place in such a volume as this.

Before proceeding further it may be well to consider certain outstanding differences between the situation in high school and that in elementary school with regard to homogeneous or ability grouping. Although there are some exceptions both ways, it is generally true that in the elementary school pupils carry their subjects for several years each, whereas in the high school they pursue them for much shorter periods of time. Except for English and in some cases a foreign language, pupils rarely carry any one high-school subject for more than two years, and in many cases for not more than one year or even one semester. This is in part a disadvantage to the high school in that certain provisions and adaptations suitable to the elementary school cannot well be

used therein. For example, it is very common in elementary schools to allow bright pupils to skip a semester or even a year's work. Since most subjects are pursued through a number of years and to a considerable extent each year's work involves a review of what has come previously, pupils bright enough to deserve being skipped ahead can make up the loss. In high school, however, no matter how bright a pupil is and how good a record he makes in his first semester's work in most subjects, he will probably have little opportunity later to learn the content of the second semester's work in those subjects if he is allowed to skip it. Also long time plans running through several years for providing for groups of differing abilities may be employed in elementary school, but not in most high-school subjects. Another disadvantage is that in the ordinary elementary school it is usually reasonably satisfactory for pupils to be in the same ability group for all subjects, that is, the bright or superior pupils are in the highest group for reading, arithmetic, spelling, history, and all other subjects in which there is any grouping. In high school this is much less satisfactory, indeed often practically impossible because of the exigencies of arranging the schedule. Despite these disadvantages, however, many high schools have made use of one plan or another of homogeneous grouping or other provisions for individual differences.

Advantages and disadvantages of homogeneous grouping.[1] During the past few years [2] homogeneous or ability grouping has been one of the most debated topics in education. Many questions have been raised concerning it, many arguments advanced both for and against it, many experiments carried on in the attempt to determine its value, many plans of diverse sorts adopted, and much attention of other sorts given it. Despite all this, most of

[1] There has frequently been some confusion as to the distinction between homogeneous or ability grouping and individual instruction. Homogeneous grouping is intended to help in providing for individual differences, but is in no sense purely individual instruction. It may be said to represent the half-way point between the usual type of class organization in which no effort at all is made to group pupils according to their abilities and the completely individualized form. Many of the arguments and statements about one apply to the other, but many do not.

[2] Homogeneous grouping is much older than standardized tests, but it is only since standardized tests have become available that it has occupied a consistently prominent place in educational thought and discussion.

the important questions raised concerning it cannot yet be answered authoritatively. Most of the arguments in favor of homogeneous grouping may be united into one, that it makes for more efficient learning and, as a result, better achievement, on the part of pupils. In most cases the basis of this argument is theoretical and not experimental. The chief reason why learning is more efficient is claimed to be that conditions resulting from homogeneous grouping, especially that a single instructional group of pupils does not include such a wide range of ability as the ordinary class group, must necessarily lead to this result.

In addition to the theoretical arguments, however, there have been many experimental efforts to discover whether or not they are true and many studies of results from various plans of ability grouping. Results from these vary so widely that it is difficult to summarize them adequately and yet briefly. Those on both sides of the controversy have made use of the results from different studies, and even in a few cases of the results from the same studies, to support their contentions. The writer believes, however, that it is a fair presentation of the case to summarize the results of such investigations as follows: with homogeneous grouping superior pupils usually do either more or better work or both than pupils of equal ability in ordinary or undivided classes; average pupils do about the same quantity and quality of work; inferior pupils on the whole probably advance somewhat more slowly, but do a better quality of work and fail less often. In almost every one of the investigations that have yielded results apparently unfavorable to ability grouping, and of which the writer has seen a detailed report, he is able to pick out some one or more factors that in his opinion were avoidable, and that probably accounted for the unfavorable results. For example, in some cases many, perhaps even almost all, of the teachers in charge of the groups differentiated on the basis of ability were unfavorable to the plan. In others an attempt was made to have the superior groups go too rapidly or cover too much additional work, perhaps also to have the average groups likewise do too much, and the inferior groups as much as average pupils should be expected to do. Sometimes the method of organizing and handling the scheme has been such as to attach a definite stigma to the inferior groups, and perhaps to cause a feeling of undesirable superiority

on the part of pupils in the superior ones. Sometimes the work has been organized and started without proper preparation and the teachers left to carry it on as they wished, with little or no direction or help. The reader can easily see that any one of these as well as of many other possible conditions would tend to exert a strong effect against the success of homogeneous grouping and to prejudice the results in favor of the ordinary form of class. In view of this the writer believes the statement justified that the statistical results so far collected and published show that homogeneous grouping may have a distinctly desirable effect, but that in order to insure this it must be properly planned and supervised.

Another important argument for ability grouping is that it makes the work of teachers easier. It is probably unsafe to press this too far, certainly as far as applying it to individual instruction. On the other hand, there is little doubt that, other things being equal, it is easier to handle a relatively homogeneous than a relatively heterogeneous group of pupils. Evidence shows that in the ordinary ungrouped class the brightest pupils have at least from two to five times as much knowledge of the subject as the dullest, and that there is a great deal of overlapping of the superior pupils of one class with the inferior ones of more advanced classes in the same subject. This range of ability and overlapping is considerably reduced by homogeneous grouping, and thus teachers are not faced with the problem of providing for individuals of such extremely different capacities and achievements in the same groups.

A third argument that perhaps deserves separate mention, although it may be thought of as a part of the first, is that of the effect upon pupils' attitudes. Although there is some difference of opinion about this point, the weight of evidence seems to be that if pupils realize that the attempt is being made to adapt the work to their capacities, their attitude toward school work will be considerably improved. Superior pupils are less liable to form habits of mental slothfulness, but more likely to exert themselves to do somewhere near their best. Neither are inferior pupils so liable to acquire the habit of failure, but instead deal with tasks of which they can do a considerable portion successfully, and thus get the valuable habit of success.

Among the arguments put forth in opposition to ability grouping the one that has received the most publicity is probably that advanced by Bagley [8] and others who charge that the scheme is undemocratic and label it "educational determinism." In other words, they claim that such a plan assigns pupils to educational groups or levels at which they must remain through their school careers, and thus does not hold open to them the equal opportunities that it is assumed every inhabitant of a democracy should have. There are several answers to this, two of which, at least, are fundamental. One is that if the proper basis of classification is used, pupils are grouped according to their known mental capacities and achievements. What they are able to do is not the result of ability grouping and the testing connected therewith, but already exists and is merely discovered thereby. The situation is similar to that in measuring height, for example. No one believes that a pupil's height is in any way affected if he is measured and for purposes of physical education or something else assigned to a group of pupils of about his height. Likewise measuring a pupil's intelligence and other mental traits, his achievement in school subjects, and so forth, and classifying on the basis of what is learned about them, is merely a recognition of what already exists. It should also be noted in this connection that regardless of what basis of classification is used the groups so formed should be flexible, that is, it should be comparatively easy to transfer a pupil from one group to another if it becomes apparent that his original placement was erroneous. Thus entirely apart from his original classification a pupil determines the group in which he belongs by the quality and quantity of his mental achievement. The second fundamental answer to this argument has to do with the true meaning of democracy. It is not, as some of the opponents of homogeneous grouping seem to assume, that each individual should have identical opportunity with every other person, but rather that he should have the opportunity by which he is able to profit most. This is just what is attempted in homogeneous grouping and individual instruction. Just as it is recognized that

[8] Bagley, W. C. "Educational Determinism; or Democracy and the I.Q." *School and Society*, 15: 373–84, April 8, 1922. Also in: *Educational Administration and Supervision*, 8:257–72, May, 1922.

———. *Determinism in Education*. Baltimore: Warwick and York, 1925. 194 p.

all children will not enjoy the best possible health as a result of the same exercise or the same diet, so they will not all receive the maximum possible mental development from studying the same content in the same way.

Another one of the arguments against homogeneous grouping is that it tends to cause too great a feeling of superiority and self-conceit on the part of the superior pupils and of discouragement on the part of the inferior. This has already been largely answered by the last argument given in favor of the plan. The majority of testimony from teachers and others who have been in positions to observe the operation of various plans of ability grouping closely is that when superior pupils are placed in groups composed entirely of similar pupils they have less feeling of superiority or conceit than when in an ordinary group. The reason is that the comparatively few superior pupils in an ordinary group stand out above the average and inferior pupils much more clearly than even the best of a group composed entirely of superior pupils do above the others in the group. In the latter pupils have to work harder and put forth more effort to rank relatively high and thus do not have the feeling that they are so gifted by nature as to be easily superior to almost all others. A similar condition holds for inferior pupils, that instead of being persistently discouraged by being at the bottom of the class and so evidently inferior to half or more of the pupils therein, they are able to compete on more or less equal terms with those in their own group, and thus have much greater incentive to work. It is possible by improper handling and overemphasis upon differences between groups to produce unfavorable attitudes on the part of pupils, but it is not very difficult to avoid this, and it does not appear to happen to any considerable degree in most schools that have homogeneous grouping. The claim that dull pupils are stimulated by bright pupils and learn from them, and that they will be deprived of this advantage if put in sections by themselves, has no real evidence to support it, but at least some to disprove it.

An argument sometimes raised is that if superior pupils are grouped by themselves, so much pressure will be placed upon them that they will injure their health through overwork. The answer to this is that for groups of pupils it almost never happens in practice, and that there is no necessity of its ever happening.

In general whether or not pupils overwork depends upon their individual dispositions and attitudes, and those who would overwork if placed in superior sections are very likely to do so even in ordinary sections. Possibly the danger is increased somewhat, but surely almost every experienced teacher will agree that far more than 90 per cent of the pupils enrolled in our high schools, even of those making the highest marks, could devote considerably more time to study without being overworked.

Those who favor individual instruction sometimes oppose homogeneous grouping on the ground that it is only a partial step to complete individualization, and that there is no reason why it should be taken instead of going the whole distance. In other words, they argue that no measures at all are better than half measures, especially since they believe that the adoption of a plan of homogeneous grouping may serve to delay that of individual instruction. Even if it is granted, which it is not, that individual instruction is entirely desirable, this argument does not seem valid. Instead, if it seems desirable to take a step in that direction, and homogeneous grouping is such a step, it is to be commended. The best thinkers in this field, however, do not believe that entirely individualized instruction is desirable, but rather that certain phases of work in which there is little to gain from interchange of opinion and other social procedures may well be individualized, whereas others, such as literature, history, and so forth, in which appreciation, attitude, opinion, and so on, play a large part, would lose much of their value if taught by purely individual methods.

The basis of placing and promoting pupils.—When homogeneous grouping was first introduced many years before standardized tests appeared, the basis of classifying pupils was ordinarily the same as that commonly employed for promoting or failing them, that is, according to school marks during the term just ending. In some cases this was supplemented by teachers' estimates of intelligence and perhaps occasionally by other data. It was not, however, until standardized tests of intelligence and achievement were available that any other basis began to be widely used. At present most school systems that have plans for caring for individual differences in one way or another make some use of one or both varieties of standardized tests. Sometimes these alone

determine placement, in other cases they are employed in combination with school marks or teachers' estimates. In a few cases they are not employed at all. Indeed, the diversity of practice, at least with regard to details, is very great. For example, some systems use point scores on intelligence tests or mental ages alone; others use intelligence quotients; some employ scores on a single achievement test for grouping in the subject with which the test deals; some make use of a series of tests in each subject; some use the average or composite scores on tests in several subjects, and so on to almost endless variety. Furthermore, there is great difference in the weights given the various factors when more than one is employed; sometimes standard test scores and school marks each count half, sometimes one much more than the other; sometimes intelligence test scores count half and achievement test scores half, and so on. In some systems such factors as pupils' health, their attendance records, their attitudes, and so on, are consciously taken into consideration, although usually they are not.

The one guiding principle that may be enunciated in this matter is that pupils should be placed according to their most probable future achievement in the subject or subjects concerned. Therefore whatever test scores or other data will predict future achievement most exactly should be employed. The statement of this principle, however, does not solve the problem, since in most cases we do not know what data best fulfill this requirement. Since a considerable portion of another chapter is devoted to prognosis, the matter will not be discussed in detail here. A few apparent facts may, however, be stated. In the case of pupils beginning subjects, it is ordinarily true that intelligence test scores are the best single predictive measures commonly available. If a subject, although new, is very similar to one that has been carried before, marks or test scores in the other subject may furnish a better prediction than intelligence test scores. For example, Latin marks have comparatively high predictive power for French. In the case of the continuation of subjects already carried, the results of standardized tests in those subjects and of previous school marks therein are probably most useful. If pupils continue to carry the same subjects with the same teachers or in the same departments of the same schools, and these departments have great uniformity in

their work, marks will probably predict future achievement at least as well as standardized test scores. If neither of these conditions is fulfilled, however, the test scores are generally preferable. In either case, that of a new subject or of one previously studied, a general average school mark for the past semester or year, or perhaps for a somewhat longer period of time, usually furnishes fairly good evidence as to what to expect and in some cases is better than any of the other measures just mentioned. Health need rarely be considered, although occasionally when it is rather doubtful whether a pupil should receive special promotion or not, or whether he should be placed in a superior section, certain types of poor health may be sufficient cause for not putting him ahead. There are no other factors that need receive consideration in most cases, although in exceptional instances there may be.

A second principle subordinate to the first may be stated as follows: If any one of the several most important kinds of data indicate that a pupil deserves to go ahead, whether it be an ordinary promotion, placement in a superior section, or something else, he should ordinarily be so placed. For example, even though his intelligence test rating is low, if his actual school mark in previous study of the subject concerned or his school mark in all subjects carried, if he has not had this particular one, is passing, he should usually be given the opportunity to go ahead with it. On the other hand, if his school mark is low, he should usually be allowed to go ahead if his scores upon standardized tests in the subject are high. It is somewhat doubtful if a high intelligence test score alone should justify letting a pupil go ahead in a high-school subject in which he has done distinctly failing work, but it should usually be considered sufficient evidence to let him try a new subject even though he has not done satisfactory work in others. If standardized test scores in the subject at issue are low, but school mark passing, a pupil should probably be allowed to advance, although the discrepancy may be due to faulty marking. In other words, the second principle stated may be summarized as: Give the pupil the benefit of the doubt when some of the evidence indicates that he is able to go ahead and some that he is not, unless the preponderance of the latter is decidedly great.

In a few high-school subjects, such as Latin and mathematics, there are at present fairly satisfactory prognostic tests. These

have been described in their appropriate chapters earlier in the book. Where they are available these should be employed in the case of pupils beginning new subjects, since usually they furnish the most reliable single means for predicting success therein. Usually they are not satisfactory for use when a pupil has already studied the subject for some time.

It has been more or less assumed so far, although not explicitly stated, that in a plan of homogeneous grouping there will be three groups of pupils; superior, average, and inferior. In very large schools there may well be five groups, including one for very superior and one for very inferior pupils, but in most instances in which the groups are intended to be at a definite number of capacity levels, three are sufficient. A more ideal scheme is not to have any predetermined number of groups or ability levels, but in dividing all pupils in a given course into as many sections as are provided, to arrange them in order of probable future achievement in the subject. For example, if the size of sections is set at twenty-five, the twenty-five pupils who are likely to be the twenty-five best in the subject should be in one section, the next twenty-five in another, and so on down to the twenty-five likely to do the poorest work. In most cases there are, however, rather serious difficulties of organization in the way of such a scheme. The chief of these is the arranging of schedules. Since the same pupils rarely if ever happen to be in the same sections of their various subjects, and since all sections of the same course can almost never be offered at the same hour, it is almost always impossible to carry out such a plan without modification. In the second place, from the standpoint of organizing curricula to provide different quantities or qualities of work to be done by pupils of various degrees of ability it is much easier to plan a certain number of definite courses than to have the number vary from semester to semester. From the practical rather than theoretical standpoint, therefore, it is recommended that there be three groups of pupils: superior, average, and inferior. This should not be interpreted to mean that if schools are able to carry through the other plan of grouping described above they should not do so.

Probably the next question that arises has to do with what proportion of pupils should be placed in each of the three ability groups. If there are in a given course enough pupils to make three

groups and no more, it is ordinarily best to divide them about evenly. Usually sections of superior pupils should be somewhat larger, and those of inferior ones somewhat smaller, than those of average pupils to be taught with the same amount of effort. For example, thirty superior pupils or twenty inferior ones are equal to about twenty-five average ones in this respect. From this standpoint it might be well to make the better of the three sections the largest and the poorer the smallest. On the other hand, although there are no fixed lines of demarcation or critical points between these three groups of pupils, it seems that in most cases it is best to place about the upper fourth of the pupils in superior sections, the middle half in average sections, and the lower fourth in inferior sections. As a compromise between these two conflicting facts, the suggestion made above that where there are only three sections they be of about equal size seems perhaps best.

If there are four sections, instead of three, one should ordinarily be for superior, two for average, and one for inferior pupils. If there are larger numbers of sections this ratio ordinarily should be approximately followed. There will, of course, be differences in particular groups of pupils enrolled in certain courses or schools. If, for example, most of the pupils come from better class homes and are themselves of somewhat above average intelligence, there may well be a third or more instead of a fourth in the superior sections and considerably less than one-fourth in the inferior, whereas if the opposite is true, these proportions should probably be reversed.

So far the discussion in this section has dealt almost entirely with homogeneous grouping rather than promotion or failure. The same general principle stated at the beginning of the section, however, should govern both. Whether a pupil is promoted or failed should be determined largely with regard to his future achievement and progress rather than to what he has done in the past. The application of this principle, however, differs somewhat according to whether pupils are to continue the same subjects or not. If, for example, a pupil has completed the first semester of any subject, such as algebra or Latin, in which the second semester's work depends largely upon that of the first, he should not be promoted unless it appears that what he has learned during the first is sufficient to give him a fair chance of success in the second.

The same is true at the end of a year's work in subjects, such as English and foreign language, that are continuous. In some subjects, however, this does not hold so strongly. The second semester's work in history, for example, ordinarily does not depend to so great a degree upon that of the first as is true in the case of some other subjects. A pupil who does not know the first semester of history well is not, therefore, so greatly handicapped in the second semester's work as he would be in a similar situation in algebra or Latin. In the case of marks given at the completion of a subject, such as at the end of a year of physics or chemistry, the chief consideration in deciding whether to promote or fail a pupil should be how well he has done the work of the year or of the second semester, as the case may be. In actual practice these diverse bases of determining promotion should not result in any real difference in standards for the promotion of pupils who are in different situations with regard to the continuation of their subjects.

Variations in the work of ability groups.—After ability groups have been organized the next question probably has to do with the matter of what differentiation should be made among them as to their work. Two general bases have been employed. One of these is rate of work, that is, the groups cover just the same work, but the superior groups do it more rapidly than the average and the average than the inferior. The other plan is frequently referred to as "enriching the curriculum." According to it inferior groups cover what may be called the minimum essentials of the subject; average groups cover these and some additional work and superior groups all that average groups do and still more.[4] In many cases the plans used are combinations of these two elements in varying proportions.

There has been considerable argument as to which one of the two plans is the better. In general the arguments advanced reduce themselves to the question of whether or not it is desirable that

[4] The statement in the text above does not mean that average groups do every particular exercise done by inferior groups and superior groups every one by average groups, but that they cover in each case all the topics of the lower group. Especially in the case of practice or drill material a higher group frequently needs fewer exercises of a particular type in order to learn the point they are intended to illustrate, and thus may often well omit some of the particular bits of work done by a lower group.

pupils save time and thus complete a given unit of school at an earlier than normal or common age. On the one side it is urged that this is a desirable saving of time and that the younger pupils are at the completion of any given stage in the educational process, the more likely they are to continue their education and thus progress further before they finally cease school attendance. On the other side it is argued that if pupils are allowed to gain time thus they will be too immature both to associate with most pupils in their grades and to leave school at the completion of high-school or any other particular time. It seems to the writer that there is some truth in the contentions of both parties and that the best procedure is to compromise between the two extremes. In other words, he believes superior pupils should be allowed to gain some time during their elementary and high-school careers and that they should also have some enrichment of the work they carry. He would suggest as limits to the time gained two years in the elementary school and one year in the high school, admitting, however, that there may be occasional very exceptional pupils who may well be allowed to gain more. For most pupils who are able to gain time, he would say that it should not be more than one year in the elementary school and one semester in the high school.

There are, however, some very real difficulties in the way of allowing pupils to gain time in particular subjects in high school. Suppose, for example, that ability groups are formed in freshman algebra. If the superior group includes, as was recommended above, about the upper one-fourth of the pupils and the total group of pupils is approximately normal, it is probable that this superior group can do the regular work of the year in about 80 or 85 per cent of the time required by average pupils. In other words, if the school year is nine months, they can probably complete the work in about seven and one-half, or if it is ten months, in eight or eight and one-half. The question then arises as to what they are to do during the remaining month or two. Probably the easiest solution is to let them go ahead with part of third semester algebra. If this is done, they will complete a portion, perhaps about half, of that semester's work and thus during the year have completed about two and one-half semesters' work in the subject. They can, of course, be given this much credit and in a

sense the problem of what to do with them is solved. However, the giving of credits in this manner is liable to cause more or less trouble later, especially in the case of pupils who wish to enter institutions of higher learning. Many of these require that for a semester's credit a pupil pursue a subject forty or forty-five minutes a day, five days a week, for sixteen or eighteen weeks, and object to giving more than two semesters' credits on a year's work no matter how much was covered. Also there is difficulty from the standpoint of what to do for those pupils who wish at some later time to complete the third semester of algebra. Presumably they can do so in about half of a semester and they will, therefore, have their time for the other half not taken care of unless some special arrangement is made.

Another possibility in the situation just described is that when the superior section completes the regular work of the first year, it should do nothing more with that subject for that year. This procedure, however, results in an accumulation of extra time toward the end of the year for which the pupils have little use and therefore tends to cause dissatisfaction, disciplinary troubles, and so forth. A still different possibility is that the superior section spend the whole year in covering the regular work, but by virtue of their superior ability spend less time per day in doing the work than do the other sections. This means ordinarily that there is little or no benefit gained from belonging to the superior section except that of greater efficiency in learning because the group is more homogeneous. It is easy to see that all these difficulties are avoided, although some others may be introduced, if the superior section takes all year to cover the same general content as do the other sections, but from time to time takes up enough additional topics and enriched subject matter to make the amount of time required as great as in the lower sections. A troublesome difficulty that arises with this plan, however, is that of providing satisfactory incentives so that pupils will desire or even be willing to be in superior sections when they know that membership therein entails a greater amount of work. Unless rewards of some sort such as extra credits, saving of time, and so on, are provided, it is frequently difficult to motivate superior sections at all satisfactorily.

Probably the best solution of the difficulty in most cases is that

mentioned above of allowing all three groups of pupils to cover a body of content within the same period of time and allowing superior pupils to profit by the fact that to do this will not require so much of their time as of other pupils' time by carrying one or occasionally more extra subjects. They will thus be enabled to complete the high-school course in somewhat less than the regular time and will also derive learning benefits from being in relatively homogeneous groups. A small amount of additional material may be introduced for average and superior groups in this plan without making them feel that they are being given additional work with no profit to themselves, especially if it is made clear to them that they are not required to do as many exercises of certain types as are the inferior sections. This is not recommended as being the most desirable procedure theoretically, but as being that which, in the situation as it actually exists, will most often obtain many benefits of homogeneous grouping and at the same time be not overly difficult to organize and administer. Likewise in such a plan the inferior pupils who supposedly spend more time per day upon their work in each subject than either the average or the superior may compensate for this by carrying less than the normal amount of work. It is only fair to say, however, that the available evidence on this point does not prove that the quality of work done by inferior pupils is appreciably improved if they reduce the number of subjects carried. The writer believes, however, that if proper attention is given the matter and the proper instructional methods are employed in the inferior sections, it will ordinarily be beneficial for inferior pupils to carry not more than three full-time subjects instead of the usual four, perhaps adding to these one or two part-time subjects not equivalent to another full-time one.

It is not only important to the success of a plan of ability grouping that there be some differentiation in the rate or amount of work done by the different groups but also that the teaching methods employed differ somewhat. It is through the adaptation of methods to more homogeneous groups that a considerable part of the increased efficiency possible from the plan may be obtained. It has already been suggested that the lower groups need more drill and practice material than the higher. This implies that in the method of handling material more of what may be called the

"drill method" may be used in inferior than in average, and more in average than in superior, groups. There should likewise be differences in the types of questions asked, more reasoning and thought questions being employed in the higher than in the lower groups, and more purely informational and factual ones in the latter than in the former. Incentives and methods of motivation should also vary. Probably more conscious attention will need to be given to them in inferior than in superior groups, since the latter will be motivated more through the urge for mental activity. Similarly in many other points differences should be made and methods as well as subject matter adapted to the pupils being dealt with.

From the standpoint of supervisors and administrators there are, at least in the larger schools, such matters to be considered as the assignment of certain teachers to superior groups and of others to average and still others to inferior groups, according to which they seem best suited to handle. It is, for example, especially important that the teachers of superior groups be themselves intellectually superior, and also that the teachers of inferior groups be unusually patient and able to give clear explanations. If, however, teachers do not appear to have characteristics that make them particularly suited for one group rather than another, it is probably best that each have groups on at least two levels, and perhaps on all three.

Miscellaneous suggestions on homogeneous grouping.—Very little has been said so far in this chapter on the matter of placing pupils by subjects, but instead the discussion has been confined chiefly to placing them within a given subject. The other phase of placement is also important, especially so, of course, in the case of elective subjects. The discussion of it in detail will, however, be omitted here and included in Chapter XXII, since it seems rather to belong there.

Two or three references have already been made to one very real hindrance to the organization of ability groups in many schools, that is, the difficulty of arranging the schedule to accommodate them. In the small school in which there is only one teacher of a subject it will, of course, be impossible, even if there are enough pupils to do so, to have more than one ordinary ability group meeting at once, and, therefore, any assignment of pupils to

groups according to their ability will have to be modified so that they can carry their other subjects. There is, however, a possible way of providing homogeneous grouping in such situations. This is grouping within the single class or section rather than grouping embracing several sections. For example, if a teacher has thirty pupils in a class, she may place the superior ones, perhaps seven or eight in number, in one group, about as many inferior in another, and the remainder in a third.

Many teachers will perhaps at first thought object to the plan just stated as being too difficult to handle within a single room and an ordinary class period. Experience with it, however, shows that this objection is not justified. The superior pupils require very little time, probably on the average little if any more than five minutes per day, and the rest of the time can be divided between the average and the inferior groups, the latter ordinarily getting more. The three groups should be handled in a sense as three separate classes, the teacher conducting their recitations in turn, the other two groups studying while each one recites. It is probably best to make the circuit twice, the first time merely inquiring about difficulties and assigning any work to be put on the board or done in class, and the second time conducting the regular recitation. The writer has known several teachers who have conducted work in this manner and who found it easily possible to give each of the three groups the time needed within a forty or forty-five minute period. This plan, however, is not recommended for use in all or perhaps even most high-school classes. As was stated earlier in the chapter, there are certain subjects and portions of subjects in which individual instruction does not appear to be satisfactory because a considerable part of the value to be gained from their study comes from class discussion and the interchange of opinions and ideas. In these subjects this plan, although preferable to individual instruction, scarcely provides sufficient time for discussion purposes. On the other hand, in those subjects such as mathematics, first-year foreign language, most laboratory work, and so on, in which there is little or nothing to be gained from interchange of opinions, and in which, therefore, individual instruction is satisfactory, this plan also may well be employed. Likewise it is suitable for certain subjects in which there are some perhaps rather minor phases that seem to deserve general discus-

sion, since these can either be taken care of in the time available for the different groups or, if desired, the three groups can unite to discuss them.

A question that has not been considered in this chapter, but that seems to deserve mention, is that of the desirability of what might be called a double system of marking and placement in high school, one not leading to graduation or perhaps to graduation but not to college entrance, and the other to graduation and college entrance if desired. According to this it is provided that pupils be promoted or failed and otherwise placed in their subjects differently in view of their future intentions, there being a higher standard for the latter of the two groups described above than for the former. Such a plan is to some extent in agreement with the suggestion made above that pupils' promotion or failure should be more or less determined by how well they will probably do in their later work, and thus pupils who, although able to do passing work in their high-school subjects would find it decidedly difficult to do so in college, should not be passed if so doing leads to college entrance, but should be if it does not involve this. In most schools, however, in which a differentiation is made on this basis, it is not a matter of passing and failing merely, but either of choice of subjects, certain combinations of subjects leading to a college-entrance diploma, and others not doing so, or of some such requirement as that pupils must earn marks of 10 per cent or one letter above passing to be recommended for college entrance, although not for high-school graduation. The writer believes that some differentiation of this sort is desirable, and that numerous pupils who cannot do successful work in the ordinary collegiate institution should be given passing marks and allowed to graduate from high school. They may then be able to continue their education in special types of schools, but not in regular academic or professional courses.

Summary.—Despite the fact that in comparison with the elementary school the high school suffers certain disadvantages in providing ability groups or individualized instruction, it has manifested much activity along this line. Although a number of arguments have been advanced against homogeneous grouping, most of them are either based on incorrect assumptions or deal with abuses rather than necessary consequences of the plan. More

efficient learning by pupils and lessening of the work of teachers appear to result from it. Experimental evidence differs somewhat and often is not easy to interpret, but seems to indicate that beneficial results can be secured. The aim should be to place pupils so that their future achievement will be as satisfactory as possible. Many bases of placing pupils have been employed, but in most of those now used standardized tests are important factors. In most schools there should be three kinds of groups, superior, average and inferior, with some differentiation in work covered and teaching methods employed. The former may be in rate or amount; a combination of both is perhaps better than either alone. Sometimes, especially when difficulties in schedule-making or school organization hinder the formation of the usual type of ability groups, the making of groups within a single section is desirable. Promotion and failure, whether connected with homogeneous grouping or not, should be determined largely by their effect on pupils' futures.

BIBLIOGRAPHY

Ashbaugh, E. J. "Homogeneous or Non-Homogeneous Grouping," *Journal of Educational Research*, 9:241–45, March, 1924.

Briggs, T. H. "Provisions for Abilities by Means of Homogeneous Grouping," *Third Yearbook of the National Association of Secondary-School Principals*. Menasha, Wisconsin: George Banta Publishing Company, 1920, p. 53–62.

Cook, R. R. "A Study of the Results of Homogeneous Grouping of Abilities in High-School Classes," *Twenty-Third Yearbook of the National Society for the Study of Education*, Part I. Bloomington, Illinois: Public School Publishing Company, 1924, p. 302–12.

Corning, H. M. "Ability Grouping in the High School," *After Testing—What?* Chicago: Scott, Foresman and Company, 1926, Chapter VI.

Feingold, G. A. "The Sectioning of High-School Classes on the Basis of Intelligence," *Educational Administration and Supervision*, 9:399–415, October, 1923.

Freeman, F. N. "Sorting the Students," *Educational Review*, 68:169–74, November, 1924.

Goble, W. L., et al. "Report on Homogeneous Grouping in Illinois High Schools," *High School Conference Proceedings*, 1923. Urbana: University of Illinois, 1924, p. 55–67.

Hughes, W. H. "Provisions for Individual Differences in High School Organization and Administration," *Journal of Educational Research*, 5:62–71, January, 1922.

Johnson, F. W. "Classification of Pupils According to Ability," *Contribu-*

tions to Education, Vol. I. Yonkers, New York: World Book Company, 1924, p. 214–220.

McCall, W. A. and Bixler, H. H. *How to Classify Pupils*. New York: Bureau of Publications, Teachers College, Columbia University, 1928. 83 p.

Odell, C. W. "An Annotated Bibliography Dealing with the Classification and Instruction of Pupils to Provide for Individual Differences," *University of Illinois Bulletin*, Vol. 21, No. 12, Bureau of Educational Research Bulletin No. 16. Urbana: University of Illinois, 1923. 50 p.

Omans, A. C. "Provision for Ability Grouping in Junior and Senior High School," *American School Board Journal*, 65:55–58, October, 1922.

Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction*. Yonkers, New York: World Book Company, 1927, p. 28–42.

Stetson, P. C. "Homogeneous Grouping in the First Year of a Five-Year High School," *School Review*, 29:351–65, May, 1921.

Symonds, P. M. "Promotion *and* Ability Grouping," *Measurement in Secondary Education*. New York: The Macmillan Company, 1927, Chapters XXII and XXIII.

Washburne, C. W., et al. "Statistical Results of Experiments with Individualization," *Twenty-Fourth Yearbook of the National Society for the Study of Education*, Part II, Section III. Bloomington, Illinois: Public School Publishing Company, 1925, p. 133–221.

CHAPTER XXII

PROGNOSIS AND GUIDANCE

Introduction.—Although portions of other chapters, especially XIV and XXI, have been concerned with predicting the probable achievement or success of pupils in their school work, the problem of prognosis and guidance as a whole deserves a chapter by itself. Prognosis and guidance are sometimes dealt with separately, but their interrelation is so close that this seems undesirable. The one chief purpose of prognosis is to advise and guide pupils more effectively, and unless it contributes to this end it is not worth the labor entailed. In other words, the prognosis of pupils' probable future achievement in school or success in a vocation is a phase, and a decidedly essential one, of a complete guidance program.

Another division sometimes made is that between educational and vocational guidance. This also is liable to lead to a false separation of two things that should be dealt with as one, or at least as two closely related phases of the same subject. It is true that educational guidance is concerned chiefly with helping pupils in their choice of subjects and courses to be carried in school, and vocational guidance with helping them in selecting the occupations they wish to enter, but choice of school subjects is largely determined by vocational intentions. For example, if a boy decides to become an engineer, it goes without saying that he will carry those school subjects which fit him for such work. Likewise, if a girl decides to become a dietician she will naturally carry courses that prepare her to do so. Although it is true in the examples just mentioned and many others that most of the specific necessary courses are offered in the college or university rather than in the high school, yet there are certain high-school subjects such as mathematics and physics in the first case, and cooking and chemistry in the second, that should be and almost always are carried in high school. The choice of certain other vocations frequently en-

tered with little or no preparation above the high school, such as farming or stenography, ordinarily means that the persons so choosing carry high-school courses that give them a considerable part, if not all, of their school preparation for these occupations. Thus it is apparent that educational and vocational guidance are mutually dependent, especially the former upon the latter.

The need for prognosis and guidance.—Although practically all progressive educators as well as many persons not at all directly connected with the schools see the need for as scientific prognosis and as helpful guidance as possible, there are some who do not, hence it may be well to state very briefly why it is needed. It is a well-known fact that a large per cent of the pupils who enter high school do not remain long enough to be graduated. Figures vary for different schools and different sections of the country, but in almost none do so many as half of those who enter as freshmen complete their senior work, and in many cases the fraction doing so is not much if any greater than one-fourth. The greatest loss is during the first year. On the average probably about one-fourth of those who enter as freshmen do not begin the sophomore year. Essentially the same situation exists in institutions of higher learning as in high schools. Although there are many causes that contribute to this condition, the outstanding one is undoubtedly the maladjustment of the school to the individual. Most of those who drop out do so either because they find the work that they endeavor to carry too difficult or because they lack interest therein. By determining individuals' aptitudes and interests and as a consequence guiding them into the work appropriate thereto, which, of course, implies that such work will be offered, a large share of the present elimination can be prevented.[1]

Much the same condition exists with regard to choosing and entering occupations. It has been shown by a number of studies that a majority of individuals are very unstable with regard to their

[1] The writer does not wish to be understood as believing, especially in the case of higher institutions, that every institution should accept every candidate for entrance, and provide appropriate work for him. It is, he thinks, perfectly fitting that particular institutions should attempt to meet the needs of groups of individuals selected on the basis of their abilities, interests, or other pertinent characteristics. The whole educational system, however, particularly that financed directly by public taxation, should offer work suited to the needs of each individual.

vocational intentions. Many pupils have no fixed intentions when they enter high school, but of those that do, probably at least half change them before graduation if they remain during the full four-year period. Of those who attend higher institutions an additional large per cent likewise change their minds. Even after entering industry or business there are numerous shifts. Undoubtedly some of this is unavoidable, since even the best that the school can do is not sufficient to give satisfactory fore-knowledge of all vocations and also because it is not always possible for an individual to secure a position in the line of work he wishes to enter. A great deal of the change and resulting lack of economy, however, is due to the fact that most persons are very poorly informed concerning the requirements and possibilities of occupations. They do not know important facts concerning their preferred vocations, which when discovered later make them dissatisfied therein, and would have prevented such a choice if known sooner. Neither do they know those facts concerning other vocations which, if they were known, would lead them to prefer those vocations to the ones actually selected.

The bases of prediction.—Chapter XXI contains sufficient discussion concerning the bases of predicting pupils' probable school marks in various subjects so as to determine whether or not they should carry those subjects and if so, in what ability sections they should be placed, that little more will be added here on that point. For the purpose of determining pupils' abilities to carry various subjects the most helpful means available are a number of prognostic tests, of which practically all of real merit have been described in the appropriate chapters of this book. The chief of these are the Iowa Placement Examinations in Chemistry, English, Mathematics, Modern Foreign Language, and Physics; the Van Wagenen Reading Scales in American History, English Literature, and General Science; the Orleans Prognosis Tests in Algebra and Geometry; the Luria-Orleans Modern Language Prognosis Test and the Orleans-Solomon Test in Latin. In addition to these, general intelligence tests, most of the tests listed in the section of Chapter XIV dealing with mechanical and engineering tests, many of the tests and rating scales mentioned in Chapter XVI, and various other measuring instruments already referred to are of considerable value for prognosis. Almost all of

these, however, are devised for the purpose of predicting pupils' capacities for particular subjects and concern themselves slightly if at all with their interests therein. How well pupils actually do, however, is determined by a combination of these two elements; therefore in addition to tests and other measuring devices of the sorts just referred to there is a large place for the use of information concerning pupils' interests and desires. Some of the instruments mentioned in Chapter XVI are of some value in this respect, but it is rather through the use of such blanks and score cards as are described near the end of this chapter, supplemented by the insight into each individual pupil's capacities and mental set gained only by a skilled guidance expert or personnel worker, that such information must be acquired. In other words, as complete information as practicable concerning each individual should be assembled and upon the basis of this advice should be given. It should not be overlooked that this information should include not only measures of various mental abilities and attitudes, but in addition of physical qualities, character and personality traits, and so forth. More will be said concerning this phase of the problem somewhat later.

The determination of success.—The question of how to determine an individual's success either in school or in an occupation is an important one in connection with prognosis and guidance. For success in school the chief basis employed has been school marks to some extent supplemented by scores on standardized achievement tests. The only other basis used frequently enough to be worth mentioning is that of the time pupils remain in school. Probably a combination of these two criteria of success is better than either one alone. The pupil who remains in school the longest and at the same time makes the highest marks is probably, in so far as the school itself is concerned, getting the most out of it or, in other words, succeeding best, whereas the one who remains the shortest amount of time and receives the lowest marks is having the least degree of success. Since measures of either kind can be determined with relative ease, it is not very difficult to determine with at least a fairly high degree of validity how well pupils succeed in school, either in general or in particular subjects.

The question of success in a vocation, however, is not susceptible of such easy measurement or determination. A number of ways of

doing so have been suggested, of which several have been used in various investigations and discussions. Possibly the one that would come first to the mind of the ordinary person is that success is measured by salary received. To some extent this is true, especially within single occupations. It is very likely that a doctor who earns ten thousand dollars a year is more successful than one who earns four thousand, and that a clerk who can command a salary of one hundred fifty dollars a month is more successful than one who can secure only one hundred dollars. It is very unsafe to assume, however, that a doctor who earns ten thousand a year is more successful than a merchant who makes eight thousand, or, reversing the situation, that a merchant who makes ten thousand is more successful than a doctor who makes eight thousand. The possible remuneration to be secured in many occupations is limited in comparison with certain others, so that no matter how much ability an individual displays in some one of the first group, he can never earn nearly as much as a person of corresponding ability in one of the second group. For example, a carpenter or plasterer, no matter how able, will never receive as high remuneration as do the most able lawyers or physicians, nor can the most able of the latter ever equal a "captain of industry," a "merchant prince" or a large scale banker of the highest ability. There is no conceivable way, as society is now organized, by which a carpenter, plasterer, lawyer or physician can, by the exercise of his occupation, earn as large an amount of money as have Rockefeller, Ford, Morgan and other multi-millionaires. On the other hand, we have no way of knowing how much a given individual who becomes a physician or a carpenter, for example, could have made if he had entered the oil business, the automobile industry, or banking.

In addition to the fact that except within the same occupation or perhaps occupations in which the financial returns are very similar, such as plastering and bricklaying or carpentering and painting, the amount earned is not a valid measure of success, it is also true that it is frequently very difficult to determine it. Individuals are often reluctant to report their incomes or if they do report them are not truthful, and there are frequently no other practicable means of securing the information.

A second possible basis of determining vocational success is that

of fame or reputation acquired. Attempts have been made to determine the average ranking of individuals by others engaged in the same occupation or by those who for some other reason are supposedly well qualified to judge. The results have not usually been very satisfactory. In many cases, moreover, the method is not applicable since the persons whose success it is desired to measure are not sufficiently well known by a sufficiently large number of raters that reliable results can be secured. In general it is only in the case of a relatively few persons who are comparatively widely known that this method may be considered fairly valid. This limitation rules out its use in the case of many vocations and for most individuals in all occupations.

A third possible method is that of determining the quantity and quality of work done by individuals. It can readily be seen that this cannot be applied satisfactorily in the case of many types of work, but only with those in which the resulting product is such that it can be counted or measured. The ordinary factory employee, the mail clerk, the plasterer, painter, or paper hanger, the stenographer, and many others produce work that can be fairly well measured or rated both for quantity and quality. In the case of others who work under more or less close supervision, such as clerks in the ordinary store, members of a section gang under a boss, and so on, employers' or supervisors' ratings as to general efficiency may be used. These are, of course, subject to the unreliability discussed in connection with the rating of pupils and teachers in Chapters XVI and XVII, but, if proper precautions are followed, are of some value. They should, however, not be employed where more objective measures of the individual's work are available.

Two other suggested bases that are perhaps the most satisfactory theoretically but most unsatisfactory practically are those of the individual's contribution to society and of the amount of self-satisfaction secured through vocational activity. Where the differences in contributions to society are great, they may frequently be differentiated, but in most cases this condition does not hold. Especially when individuals in different vocations are being compared is there little possibility of making such distinctions. There would probably be little disagreement with the statement that the doctor, minister, or teacher who is at all worthy of the name contributes

more to society than does the day laborer. Similarily there is probably no one who would question the statement that the physician who treats, with superior skill, twice as many patients as another physician is making a greater social contribution than the latter, perhaps provided that the patients of the two physicians constitute similar groups. There is no general agreement, however, as to whether a physician makes a greater contribution to society than does a teacher of the same degree of ability or whether the work of a carpenter is more valuable than that of an equally good painter. The other basis, that of the amount of self-satisfaction derived from vocational activity, is even less usable from the practical standpoint. An individual may estimate his own self-satisfaction, but he cannot compare it reliably with that of others nor with what he might obtain in another calling. Others cannot make reliable enough estimates to justify their use.

Probably the best that can be done in the situation as it now exists is to attempt to measure vocational success in one of two ways, or else to be satisfied with a very general and unreliable estimate thereof. The two ways referred to are by salary received and by amount or quality of work. As stated above, however, both of these are almost entirely limited in their usefulness to the comparison of different individuals in the same or very similar occupations. The other method referred to is that of using the combined opinions of a number of supposedly competent judges who endeavor to take into account all known factors as to the vocational success of various individuals. Neither of these is even approximately as satisfactory a criterion of success in vocational activity as are school marks, test scores and length of time in school for measuring success in school.

Because of this lack of satisfactory criteria of vocational success the procedure in connection with vocational guidance has in a certain respect differed considerably from that in educational guidance. The difference has been that apparently successful individuals in various vocations have been selected and efforts made to determine their abilities and traits in the belief that others possessing similar ones would probably also be successful in the same vocations. Although this procedure is liable to one serious fallacy, that individuals successful in a given vocation may happen to possess certain traits that not only do not contribute to success

therein, but may even hinder it, or that would cause even greater success in some other vocation, yet it has yielded considerable helpful information.

In this connection the erroneous use that some persons have attempted to make of certain intelligence test scores obtained from the army testing program during the World War should be mentioned. Among the many data published as a result of this testing are distributions of intelligence-test scores of men according to the vocations they had followed before entering the army, or rather according to the ones they claimed to have followed. These results show that individuals claiming to have been engaged in such occupations as engineering, the ministry, medicine, and accountancy ranked very high, whereas others, such as fishermen, laborers, textile workers and cooks were low, and many others such as artists, druggists and gunsmiths, in between. These statements, of course, apply to the groups as wholes, not to all individuals therein, since there was a great deal of overlapping between the groups. Some persons interested in guidance have made use of these figures by stating that their central tendency, usually the median, or perhaps the range between the first and third quartiles, that is, the middle half of the scores, indicate the intelligence that should be possessed by individuals planning to enter the various vocations.

There are several reasons why the deduction just stated cannot properly be drawn from these data. One has already been hinted at, that the scores were tabulated according to reported rather than actually known occupations. There was undoubtedly a tendency for many men to report themselves as belonging to occupations standing somewhat higher in general social esteem than the ones to which they actually did belong, or perhaps to occupations to which they hoped to be assigned for army work rather than their true ones. Also the mere fact that individuals of a certain intelligence level happen to be in a particular vocation is by no means proof that they possess the degree of intelligence best suited to success therein. Still another objection to the use of these data in the way mentioned above is that in many cases the groups of men tested did not constitute random or average samplings of the total vocational groups to which they belonged. The groups of men entering the army from different vocations did not all repre-

sent the same basis of selection. In some cases they were average groups, in others they were above the average of individuals actually engaged in certain vocations, and in others below. Despite these and perhaps other facts which indicate that these data are not valid and accurate indications of the degrees of intelligence desirable in the different vocations, they do have some value in this connection and may be considered as general guides, provided they are used cautiously and intelligently.

The basis of prognosis.—The particular items of information upon which predictions of the probable quality of pupils' school work may best be made were mentioned in the last chapter. The chief items that should be included are intelligence test scores, marks in the same or similar school subjects and also general average school marks, standardized test scores in the same or similar subjects, and prognosis test scores. The point was made, however, and cannot be too much emphasized, that it is desirable to employ all available information that seems to have any bearing at all upon the matter. In connection with the use of this information, however, nothing was said concerning one point, how to combine the available facts in such a way as to give the greatest accuracy of prediction. Anyone who has given the matter a few moments' thought realizes that they are not all equally valuable for the purpose; therefore the question arises as to how to weight each, in other words, how largely to let each enter into the prediction made. To do this in the most accurate fashion requires the use of multiple or partial regression [2] equations and these in turn that each

[2] Since partial or multiple regression as well as partial and multiple correlation are among the relatively more difficult statistical procedures employed in education, and there are likewise few occasions upon which the ordinary teacher needs to employ them, they are not explained in the statistical chapter of this book. The partial or multiple regression equation may be defined briefly as an equation by means of which the most probable value of one variable may be predicted from the known values of several others with which it has been correlated. In other words, a pupil's probable mark in a school subject may be predicted from a combination of the prognostic data available. Methods of computing them may be found in any relatively advanced text on statistics and in some of the more elementary ones. For example, see

Odell, C. W. *Educational Statistics.* New York: The Century Co., 1925. p. 245–63, or

Holzinger, K. J. *Statistical Methods for Students in Education.* Boston: Ginn and Company, 1928. p. 283–315.

of the various factors employed be correlated with each of the others. For ordinary work, however, the teacher can hardly be expected to make use of the methods of partial or multiple regression, but must use more approximate ones. Probably the best substitute in most cases is merely to estimate as carefully as possible the relative weights of the different factors to be used and to combine them on that basis. For example, if an attempt is being made to predict pupils' marks in geometry on the basis of intelligence-test scores, algebra marks, and a prognostic test in geometry, and it appears that the latter gives as good prediction as the former two combined and that they are about equally good, they should each be weighted one and the prognosis test score two.

An important requirement from the practical side is that any method of prediction must be reasonably easy, quick, and accurate. Although it is possible in special investigations and studies to devote many hours to each individual pupil in the attempt to predict what he will do, this is not practicable in the ordinary school situation. What must be done there is to select the few items of the greatest value that also can be secured with comparative ease and economy and employ them. Likewise it goes almost without saying that unless the resulting predictions have a reasonable degree of accuracy they are not worth the trouble required to make them. Different persons may disagree as to how to interpret the word "reasonable" in the preceding statement, and the writer will make no attempt to define it in exact numerical terms. Since they cannot take everything, pupils have to take certain subjects rather than others, and if any means of guiding them in so doing results in appreciably greater success on their part than would result if no guidance were afforded them this means is worth using until something better is available.

One important fact revealed by studies in this field is that most of the test scores and other available items of information have greater negative than positive value. It is more likely that pupils who rate low will be unsuccessful in the school subject or vocation in question than that those who rate high will be successful. Apparently most high-school subjects require about the same degree of general mental ability and to a considerable extent the same specific abilities for success. Likewise large groups of vocations require about the same capacities. An individual who does

not possess these, therefore, can be rather definitely advised not to carry any of the given subjects or enter any of the given vocations, since he is almost sure to fail therein, but one who does possess them cannot be told definitely that he is more likely to succeed in one given subject or vocation than in another. For example, a pupil who is apparently unable to grasp grammatical principles and to apply them will have little chance of being very successful in foreign language, but this information gives no help in deciding whether he will be most successful in history, mathematics, science, or some other high-school subject. The same is even more true in regard to vocations. Anyone of decidedly low intelligence is wasting his time and energy if he attempts to become an accountant or a surgeon, for example, but there are many other vocations in which those who possess high intelligence can profitably employ it besides the two named. Likewise an individual of high intelligence who is extremely nervous and cannot control the movements of his hands can be advised that he has little chance of success as a surgeon, but not, on the basis of his nervousness alone, which one of many other occupations it would be well for him to enter. To take another example, an individual who has a very poor physique and very little strength is not likely to succeed at handling freight or other fairly severe labor, nor one who has serious trouble with his feet as a city mail carrier. Almost every occupation has certain requirements, intellectual, emotional, physical, or moral, that must be possessed by anyone to enable him to achieve success therein. Therefore if tests or other means of studying individuals show that they are lacking in any of these qualities, they should be very definitely advised not to enter the vocations that require them.

There are some exceptions to the general rule just stated and illustrated in the matter of school subjects and still more in the matter of vocations, especially as regards traits other than general intelligence. In all cases, however, the matter of attitude enters in, so that the individual who possesses all the capacities required to succeed in any given line will not do so unless he has an attitude that leads him to apply them. For example, almost exactly the same abilities needed for successful school work in French and Spanish are likewise needed in Latin, but the pupil who thinks of Latin as a dead language of no use to anyone, but expects to make

considerable use of French and Spanish in his after-school life, will probably succeed much better in the modern languages than in the ancient one.

It should not be overlooked in this connection that an individual may not only have too little capacity of a certain sort to enable him to succeed in a particular occupation, but he may also have too much. This is particularly true of general intelligence and of certain character and emotional traits. The person of superior intelligence is very unlikely to be satisfied in a position in which the work is extremely routine and monotonous and in which there is very little opportunity to employ his superior intelligence. He is, therefore, likely to become so dissatisfied in such work that it will lower the quality of what he does and thus render him less efficient and a less desirable employee than someone of considerably less intelligence who is content with his job. Even more than this there seems to be evidence to show that in some occupations undesirable, even dangerous, results follow because those of high intelligence are not content to keep their attention fixed upon the comparatively simple tasks they must perform. An individual operating a machine such as a lathe, for example, may damage the machine and the article being made, even injure himself or others, by allowing his attention to wander from what he is supposed to do. Similarly the street-car motorman who will not keep his attention fixed on the track, but is instead, as a more intelligent person is likely to do, trying to see a great deal besides what is directly ahead, is liable to have more accidents than the individual who is content to watch the track alone. The same is true of such traits or qualities as ambition, for example, since the overly ambitious person is liable to be much less satisfied in an inferior position than the one of less ambition.

Predicting success in high school and in college.—From the standpoint of predicting school success the high school is concerned with predicting the success of the elementary-school pupils who come to it, and of its own pupils who go on to higher institutions. Furthermore, as has already been stated, it is concerned with predicting both success in particular subjects and general or average success in the educational institution attended. The means of doing this have already been discussed, but little or nothing has been said as to how accurately it can be done with these means.

Among the best investigations of the reliability of predicting success in high school are those of Ross [3] and Flemming.[4] Ross' study was concerned chiefly with the correlation of high-school marks with elementary-school marks, and Flemming's with intelligence test scores and estimates of various traits. The correlations obtained varied considerably. Ross found that by properly weighting and combining marks in certain elementary-school subjects the correlation with high-school freshman marks averaged about .65 and with marks during the first two years of high school slightly lower. He found much lower correlations with intelligence, reading, and arithmetic test scores. Flemming found a correlation of about .85 between junior high-school marks and a combination of teachers' estimates of intelligence, attitude, and energy, and chronological age, and of about .78 for senior high-school marks with a combination of scores on two intelligence tests, ratings for industry and energy and chronological age. The corresponding coefficient of alienation or, in other words, the guessing element involved in predictions based on correlations of the sizes just given, is about .75 in Ross' study and about .50 for the junior high school and .65 for the senior high school in Flemming's. In other words, the element of guessing still remaining amounts to half or more.

The use of data obtained in or at the close of high school for predicting success in college gives about the same results. In a study by the writer [5] the highest correlation obtained between freshman college marks in any subject and the best possible combination of intelligence test score and marks in different high-school subjects was .63 and it was greater than .50 for less than

[3] Ross, C. C. "The Relation Between Grade School Record and High School Achievement: A Study of the Diagnostic Value of Individual Record Cards," *Teachers College, Columbia University, Contributions to Education*, No. 166. New York: Bureau of Publications, Teachers College, Columbia University, 1925. 70 p.

[4] Flemming, C. W. "A Detailed Analysis of Achievement in the High School. The Comparative Significance of Certain Mental, Physical, and Character Traits for Success," *Teachers College, Columbia University, Contributions to Education*, No. 196. New York: Bureau of Publications, Teachers College, Columbia University, 1926. 209 p.

[5] Odell, C. W. "Predicting the Scholastic Success of College Freshmen," *University of Illinois Bulletin*, Vol. 25, No. 2, Bureau of Educational Research Bulletin No. 37. Urbana: University of Illinois, 1927. 54 p.

half of the college subjects. Certain conditions in this study, however, prevented the securing of as high correlations as have been obtained in other situations and as can be obtained in many cases. Few of the reported correlations are above .80 and none with which the writer is acquainted above .90. On the whole, there is probably a tendency for the prediction of success in college to be slightly less accurate than that of success in high school.

Another tendency that appears in this connection is for the possible predictions of college success from data secured before students enter there to become lower the longer the time elapsed. Predictions of success during a whole college course based on data of the sort referred to are usually lower than those for the freshman year alone. If, however, records of the freshman year are used in connection with the other data in making predictions of later success, these may be made fully as reliable, perhaps even slightly more so, than those for the freshman year from previous data. The same situation holds with regard to high school also. It also seems to be established as a fact that although special aptitude or prognosis tests furnish the best single predictive measures of success in school subjects for those who have had little or no work in those subjects, actual achievement tests in the subjects themselves furnish the best measures after individuals have had considerable amounts of work therein.

The school's contribution in vocational guidance.—Although the previous discussion in this chapter has dealt with both educational and vocational guidance, it seems well to state in a short connected discussion what the writer conceives the contribution of the high school to vocational guidance should be. Before considering this directly it should be pointed out that in many ways the problem of the school is different from that of the employer. In the first place, the school is concerned primarily with rating capacity rather than actual ability, that is to say, with rating the possible success of individuals rather than their present ability to perform. In some cases employers are also interested in this, since they are willing to give employees the necessary training to enable them to capitalize their capacities in actual achievement, but in many other instances they are concerned with present ability rather than prospective future ability. In the second place, the school is concerned chiefly with placing the individual where he

will best fit and attain the greatest measure of success rather than in helping a business or industry by securing the best possible employees for it. Frequently there is no opposition between these two aims, since what is best for the individual seeking employment is often best for the employer. In other cases, however, there is a difference, since it is frequently to the advantage of the employer to keep in his employ at certain jobs individuals who have abilities that would enable them to perform work that is more difficult, or commands a higher remuneration or otherwise yields greater returns.

As the writer views it the duty of the school with regard to vocational guidance includes two chief phases. One is that which forms the basis of a considerable portion of the preceding discussion, that of testing the individual and securing such other information about him as will enable wise advice to be given him. The second is that of offering him opportunities for acquiring wide-spread vocational information and experience so that he himself will be better fitted to make a wise selection. The information should deal with such points as requirements for entrance, promotion and success, probable rewards, hazards and difficulties encountered, conditions of work, probable future of the vocation, demand and supply of workers, and so forth. Also there should be opportunities for actual experience in a number of different kinds of work through "trying out" courses so that pupils may gain a first-hand idea of what a number of different lines of work are like.

Because of limitations of time, however, even large high schools, much less small ones, cannot offer a sufficient amount of information in courses of the first sort or a sufficient variety of work in those of the second to cover the field at all completely. Such courses should, therefore, be supplemented by the collection in the library, personnel office, principal's office, or somewhere else, of as much useful and practical information as possible concerning various occupations. This should be available to all pupils who wish to consult it, ordinarily under the guidance of a counselor or teacher trained for the purpose. Furthermore this adviser should be able to tell pupils where they can most readily and conveniently visit businesses and industries of various sorts so that they may actually see what the work is like.

On the basis of the available test scores and the other data gathered, including the pupil's reactions to the vocational information and experiences he has acquired, some one especially trained for the work should be available to advise pupils concerning their vocational selections. Almost never if ever should a pupil be told that there is only one vocation that he should enter. Rather the advice should take the form of naming a number of vocations in any one of which it is probable that he can succeed fairly well, coupled with the endeavor by conversation and questioning to lead him to discover for himself in which one of this group his interest is greatest. Frequently it will be desirable to point out to him that there are certain vocations for which he is evidently unfitted because he lacks certain qualities necessary to success and also perhaps that there are others that he should not enter because he possesses greater capacities and abilities than they require and it would be a waste of these capacities and abilities for him to enter any one of them.

Available information blanks.—Many experienced vocational advisers have developed blanks or other means of their own for collecting in usable form information concerning pupils' vocational interests. For those who are not very well trained or experienced in such work, however, it is probably well to employ some one of the several blanks worked out by experts and made generally available. Three of these will, therefore, be described in the following pages.

Vocational Guidance Scales and Tests

J. M. Brewer (1926)

Self-Measuring Scales for Information on Education and Vocations, for Achievement and Experience in Work and Education; Vocational Information Test

The first of these instruments lists from nine to forty-four items under each of the following nine heads:

School Problems
Knowledge about Agricultural Occupations
Knowledge about Industrial Occupations
Knowledge about Commercial Occupations
Knowledge about Professional Occupations
Knowledge about Other Occupations

Knowledge of General Problems in the Occupational World
Ability to Study Occupations Intelligently
Knowledge of the Various Educational Opportunities Available

Those using the scale are to check each item about which they have fairly good information, or concerning which they feel certain, and then compute their own scores. The second self-measuring scale is very similar, including the same nine general topics, but, as its title indicates, dealing with achievement and experience rather than information or knowledge. The Vocational Information Test contains forty-three elements, mostly of the multiple-answer type, divided into the following six parts:

I. The work in various occupations
II. Educational facts about work and workers
III. Economic and social problems of workers
IV. Miscellaneous questions about workers
V. Comparing workers in pairs
VI. Classifying workers

These measuring instruments are in no sense standardized and probably never will be. Their value consists not so much in yielding a single score by which an individual can be guided, as in securing in fairly definite and concrete form facts concerning individuals useful both to them and to others in determining what vocations they should enter. Revisions of these instruments by Mildred Lincoln should be available by the time this book leaves the press.

Bureau of Vocational Guidance. 10¢ per 3; 25¢ per 12, 90¢ per 50, $1.50 per 100.

Interest Report Blank

K. M. Cowdery

This is a revision of the earlier Interest Analysis Blank published by the Bureau of Personnel Research of the Carnegie Institute of Technology. It consists of a list of about eighty-five occupations, another of almost twice as many varied interests, and finally one of about twenty-five school subjects or groups of subjects. The first and last are similar to those found in a number of other blanks, but the first half of the varied interests

list is somewhat unusual. It consists entirely of brief descriptions of different kinds of people. For example, among its items are "fat men," "fat women," "thin men," "thin women," "people with hooked noses," "blind people," "jealous people," "thrifty people," "spendthrifts," "educated people," "cautious people," "Southerners," "New Englanders," and so on. The other items in this list have to do with sports and recreations, including reading materials, characters, mostly movie actors and actresses, conditions or occurrences of ordinary life, such as "living in the city," "living in the country," "working alone," "telling a story," and so on, and a number of miscellaneous items. After each item are the letters *L, I* and *D*, standing for like, indifferent and dislike, respectively. The proper one of them is to be encircled. There is, of course, no time limit, since it is intended that all individuals will respond to all items, but it is stated that not more than thirty to thirty-five minutes should be used. Apparently this has been more or less superseded by the longer Vocational Interest Blank of Strong.

Stanford University Press.

Vocational Interest Blank
E. K. Strong (1929)

Although intended primarily for college or adult men, this appears to have value for high-school boys also. It has eight parts, of which the first lists about out hundred occupations toward each of which the individual is to indicate his like, dislike, or indifference. The second contains the names of about fifty amusements, the third of about forty school subjects, the fourth of about fifty activities, most of which are more or less vocational, and the fifth of about fifty types of people on the basis of temperament, race, habits, and various other characteristics. The responses called for are of the same type as in Part I. In Part VI are four groups: activities, factors affecting work, persons, and positions, respectively. For each group the subject is to indicate his first three choices and his last three of ten items. Part VII presents about forty pairs of expressions such as "policeman" and "fireman," "persuade others" and "order others," "physical activity" and "mental activity," "fat men" and "thin men," "nights spent at

home" and "nights away from home," and so on. The individual is to mark each pair according to whether he prefers one or the other of the expressions or likes each equally well. The last part calls for self-rating on about twenty-five items as applicable, inapplicable, or doubtful, and on about a dozen other characteristics for which three degrees of each are stated. The author suggests a time limit of thirty-five minutes, but further states that adult men require from twenty to one hundred twenty minutes, the median being thirty.

This blank is based upon earlier ones by Miner, Moore, Ream, Freyd, and Cowdery, especially the latter. The author states that this test "measures the extent to which one's interests agree or disagree with those of successful men in a given profession." Separate scoring keys have been printed for each of the following occupations:

Advertiser	Minister
Architect	Personnel manager
Certified public accountant	Physician
Chemist	Psychologist
Engineer	Purchasing agent
Farmer	Real estate salesman
Journalist (newspaper editor)	School teacher and administrator
Lawyer	Vacuum cleaner salesman
Life insurance salesman	Y.M.C.A. general secretary

A few additional ones have been developed, but are not yet printed. The differences between the scales consist in the weights assigned to responses which are such that the total score according to each scale is supposedly most valid for the given occupation. Provision is made for converting raw scores into three ratings, "A" (yes), "B" (not sure), and "C" (no), which indicate whether persons' interests in particular vocations are great enough to justify their entrance.

The validity of the scores as determined by differentiation between persons in different occupations appears to be reasonably high. Correlations between scores by individuals in more or less similar occupations run fairly high and those between occupations that have little in common are rather low. Various data on reliability have been reported. Apparently the coefficient is from .80 to .85.

Stanford University Press. 10¢ per copy, 90¢ per 10, $2.00 per 25, $3.50 per 50, $6.00 per 100, $25.00 per 500; scoring scales $1.00 per copy.

Summary.—Although sometimes separated the problems of educational and vocational guidance are so closely interrelated that they should be considered together. Both kinds of guidance are needed because many pupils either drop out of high school or do not derive maximum profit from their work therein, and likewise many individuals are either permanent vocational misfits or lose much time in finding their proper places. In order to determine what subjects high-school pupils should carry and what vocations individuals should enter, use should be made of all available useful means. Chief among these are prognosis tests, general intelligence tests, achievement tests, school marks, and information as to pupils' interests. Success in school is fairly well measured by the marks received or the time spent in school, but success in vocation is much more difficult to determine. Several means have been used and some others suggested, but none is generally satisfactory. It is, therefore, much more difficult to determine when individuals are properly placed in business or industry than in school. In using prognostic data the various items should be so weighted as to yield the best predictions possible. In using these data it should be recognized that they generally have more negative than positive value, that is, that they indicate more definitely that individuals will not succeed in certain school subjects or occupations than that they will succeed in others. Ordinarily the best that can be done on the positive side is to determine that they seem to have the capacities for success in any one of a group of subjects or vocations. Whether they actually succeed or not depends upon their interests and attitudes. The best combination of available measures for predicting success in high-school or college work gives predictions that are about half guesses and in most cases the predictions actually made are less reliable than this. The duty of the high school in the matter of vocational guidance should be to provide as much vocational information and experience as possible, to collect and interpret scores and other pertinent data, and to provide trained counselors to advise pupils. To aid in doing this many experienced counselors have worked out

their own blanks, but there are several generally available that may well be used.

BIBLIOGRAPHY

Allen, W. S. "A Study in Latin Prognosis," *Teachers College, Columbia University, Contributions to Education*, No. 135. New York: Bureau of Publications, Teachers College, Columbia University, 1923. 41 p.

Beatley, Bancroft. "The Relative Standing of Students in Secondary School, on Comprehensive Entrance Examinations, and in College," *School Review*, 30:141–47, February, 1922.

Brewer, J. M., et al. "Mental Measurement in Educational and Vocational Guidance," *Harvard Bulletin on Education*, No. 10. Cambridge, Massachusetts: Harvard University Press, 1924.

Burwell, W. R. and MacPhail, A. H. "Some Practical Results of Psychological Testing at Brown University," *School and Society*, 22:48–56, July 11, 1925.

Clem, O. M. "Detailed Factors in Latin Prognosis," *Teachers College, Columbia University, Contributions to Education*, No. 144. New York: Bureau of Publications, Teachers College, Columbia University, 1924. 52 p.

———. "Latin Prognosis: A Study of the Detailed Factors of Individual Pupils," *Journal of Educational Psychology*, 16:160–69, March, 1925.

Colvin, S. S. "Educational Guidance and Tests in College," *Journal of Applied Psychology*, 5:32–38, March, 1921.

———. "Psychological Tests at Brown University," *School and Society*, 10:27–30, July 5, 1919.

———. "The Purposes and Methods of Psychological Tests in Schools and Colleges," *Education*, 40:404–16, March, 1920.

———. "The Validity of Psychological Tests for College Entrance," *Educational Review*, 60:7–17, June, 1920.

Colvin, S. S. and MacPhail, A. H. "The Value of Psychological Tests at Brown University," *School and Society*, 16:113–22, July 29, 1922.

Edgerton, A. H., et al. "Vocational Guidance," *Twenty-Third Yearbook of the National Society for the Study of Education*, Part II, Section I. Bloomington, Illinois: Public School Publishing Company, 1924, p. 1–198.

Ernst, J. L. "Psychological Tests vs. the First Semester's Grades as a Means of Academic Prediction," *School and Society*, 18:419–20, October 6, 1923.

Flemming, C. W. "A Detailed Analysis of Achievement in the High School. The Comparative Significance of Certain Mental, Physical, and Character Traits for Success," *Teachers College, Columbia University, Contributions to Education*, No. 196. New York: Bureau of Publications, Teachers College, Columbia University, 1926. 209 p.

Fretwell, E. K. "A Study in Educational Prognosis," *Teachers College,*

Columbia University, Contributions to Education, No. 99. New York: Bureau of Publications, Teachers College, Columbia University, 1919. 55 p.

Gates, A. I. and La Salle, Jessie. "The Relative Predictive Values of Certain Intelligence and Educational Tests Together with a Study of the Effect of Educational Achievement upon Intelligence Test Scores," *Journal of Educational Psychology*, 14:517–39, December, 1923.

Gowen, J. W. and Gooch, Marjorie. "The Mental Attainments of College Students in Relation to Previous Training," *Journal of Educational Psychology*, 16:547–68, November, 1925.

Herriott, M. E. "Attitudes as Factors of Scholastic Success," *University of Illinois Bulletin*, Vol. 27, No. 2, Bureau of Educational Research Bulletin No. 47. Urbana: University of Illinois, 1929. 72 p.

Hull, C. L. *Aptitude Testing*. Yonkers, New York: World Book Company, 1928. 535 p.

————. "The Joint Yield from Teams of Tests," *Journal of Educational Psychology*, 14:396–406, October, 1923.

Hull, C. L. and Limp, C. E. "The Differentiation of the Aptitudes of an Individual by Means of Test Batteries," *Journal of Educational Psychology*, 16:73–88, February, 1925.

Johnston, J. B. "Predicting Success in College at the Time of Entrance," *School and Society*, 23:82–88, January 16, 1926.

Johnston, J. B. "Predicting Success or Failure in College at the Time of Entrance," *School and Society*, 19:772–76, June 28, 1924; 20:27–32, July 5, 1924.

————. "Tests for Ability Before College Entrance," *School and Society*, 15:345–53, April 1, 1922.

Jordan, A. M. "Student Mortality," *School and Society*, 22:821–24, December 26, 1925.

Kelley, T. L. "Educational Guidance: An Experimental Study in the Analysis and Prediction of High School Pupils," *Teachers College, Columbia University, Contributions to Education*, No. 71. New York: Bureau of Publications, Teachers College, Columbia University, 1914. 116 p.

Kemble, W. F. *Choosing Employees by Test*. New York: Engineering Magazine Company, 381 Fourth Avenue, 1917. 333 p.

McCall, W. A. "Measurement in Vocational Guidance," *How to Measure in Education*. New York: The Macmillan Company, 1922, Chapter VI.

MacPhail, A. H. *The Intelligence of College Students*. Baltimore: Warwick and York, 1924. 176 p.

Madsen, I. N. "The Contributions of Intelligence Tests to Educational Guidance in High School," *School Review*, 30:692–701, November, 1922.

May, M. A. "Predicting Academic Success," *Journal of Educational Psychology*, 14:429–40, October, 1923.

Miles, W. R. "Comparison of Elementary and High School Grades," *University of Iowa Studies in Education*, Vol. 1, No. 1. Iowa City: University of Iowa, 1910.

Odell, C. W. "Predicting the Scholastic Success of College Freshmen,"

University of Illinois Bulletin, Vol. 25, No. 2, Bureau of Educational Research Bulletin No. 37. Urbana: University of Illinois, 1927. 54 p.

Proctor, W. M. "Psychological Tests and Guidance of High School Pupils," *Journal of Educational Research Monographs*, No. 1. Bloomington, Illinois: Public School Publishing Company, 1923. 125 p.

Roberts, A. C. "Objective Measures of Intelligence in Relation to High-School and College Administration," *Educational Administration and Supervision*, 8:530–40, December, 1922.

Rogers, A. L. "The Use of Psychological Tests in the Administration of Colleges of Liberal Arts for Women," *Twenty-First Yearbook of the National Society for the Study of Education*. Bloomington, Illinois: Public School Publishing Company, 1922, Part II, Chapter IX.

Rogers, D. C. "Intelligence Examinations and College Entrance," *Smith Alumnae Quarterly*, 13:1, November, 1921.

Root, W. T. "The Freshman: Thorndike College Entrance Tests, First Semester Grades, Binet Tests," *Journal of Applied Psychology*, 7:77–92, March, 1923.

Ross, C. C. "The Relation Between Grade School Record and High School Achievement: A Study of the Diagnostic Value of Individual Record Cards," *Teachers College, Columbia University, Contributions to Education*, No. 166. New York: Bureau of Publications, Teachers College, Columbia University, 1925. 70 p.

Smith, F. C. "A Rational Basis for Determining the Fitness for College Entrance," *University of Iowa Studies in Education*, Vol. 1, No. 3. Iowa City: University of Iowa, 1910.

Smith, H. L. and Wright, W. W. "Prognosis and Special Ability Tests," *Tests and Measurements*. New York: Silver, Burdett and Company, 1928, Chapter XX.

Terman, L. M. "Intelligence Tests in Colleges and Universities," *School and Society*, 13:481–94, April 23, 1921.

———. "Intelligence Tests in Vocational and Educational Guidance," *The Intelligence of School Children*. Boston: Houghton Mifflin Company, 1919, Chapter XII.

Thorndike, E. L. "Early Interests: Their Permanence and Relation to Abilities," *School and Society*, 5:178–79, February 10, 1917.

———. "Intelligence Examinations for College Entrance," *Journal of Educational Research*, 1:329–37, May, 1920.

———. "On the New Plan of Admitting Students at Columbia University," *Journal of Educational Research*, 4:95–101, September, 1921.

Toops, H. A. "The Status of University Intelligence Tests in 1923–1924," *Journal of Educational Psychology*, 17:23–36, 110–24; January, February, 1926.

Wood, B. D. *Measurement in Higher Education*. Yonkers: World Book Company, 1923, Chapters II–V.

Yerkes, R. M. (Edited by). "Intelligence Ratings of Occupational Groups," *Memoirs of the National Academy of Sciences*, Vol. 15. Washington: Government Printing Office, 1921, Part III, Chapter XV.

Yoakum, C. S. and Yerkes, R. M. (Compiled and Edited by). *Army Mental Tests.* New York: Henry Holt and Company, 1920, p. 196–204.
"Report on the Use of Intelligence Examinations." New York: Columbia University, 1922. 27 p.

CHAPTER XXIII

DIAGNOSIS

Introduction.—The diagnosis of pupils and of their achievements constitutes the most important of the many uses of tests and test scores. The mere giving of tests is of little value, but is only a preliminary step to something really worth while. Even though tests are given for the purpose of predicting educational or vocational achievement, classifying pupils, determining promotion or failure rating the efficiency of teachers or of whole schools or systems, they may also almost always be made to yield information of at least some diagnostic value. Therefore if tests not primarily intended for diagnostic purposes have been given for any of the reasons just mentioned or for any others, teachers and others concerned should not neglect to secure such help in diagnosing pupils as may be obtained from the results. Most diagnosis and succeeding remedial instruction, however, should be based upon results from tests specifically intended for the purpose and not be more or less incidental to other uses thereof.

There can be no hard and fast line drawn between diagnostic and non-diagnostic or general tests. Practically all tests are diagnostic to at least a slight degree. Thus if a pupil takes a series of very general tests in different subjects, the scores are diagnostic of the status of his knowledge and ability to the extent that they indicate that his achievement is better in some subjects than in others. Some tests are slightly more diagnostic than this in that they indicate in a general way in which phases or divisions of a subject pupils are weak and in which they are strong. For example, an American history test divided into two parts, one of which deals with the period preceding the Civil War, and the other with that following it, will yield results which indicate that certain pupils know the first period better than the second, and others the second better than the first. To give another example, a test in

chemistry or physics that contains one part dealing with information and another with problems will yield similar information with respect to these two divisions of the subject. A test in American history that is divided to cover several periods of time instead of only two is still more diagnostic, as also is one in physics or chemistry that contains parts each devoted to information about one of the major divisions of the subject and other parts each devoted to problem solving therein. Continuing the subdivision of the whole subject and at the same time increasing the number of elements makes the resulting tests more and more diagnostic until finally a completely diagnostic test is one that tests all the items of knowledge or ability supposed to have been learned in connection with any one phase or division of a subject.

There are very few tests in high-school subjects that are completely diagnostic. Those that are so are necessarily limited in the subject matter that they cover to small portions of the whole content of any course. In many subjects it is impossible, at least from the practical if not from the theoretical standpoint, to construct completely diagnostic tests. In others or portions of them such tests can be made, but the process is often so laborious and the time required to administer the tests so great that they are not practicable. For example, in first-year foreign language a vocabulary test or series of tests can be made to measure knowledge of all the words supposed to have been learned, or a grammar test to cover all the rules of grammar. Even here, however, it is impossible to make tests so complete that they measure all possible uses or applications of the words or grammatical rules. To take another example, in the field of history, it is possible to construct a test to measure knowledge of all dates that have been studied, but not to measure whether or not pupils understand the full significance of the events occurring at those dates and all other matters connected therewith. On the other hand, it is practically impossible to test pupils' knowledge of all the detailed facts they have studied in a history course, or still more of that indefinite but important objective of history teaching often denominated "ability to think historically." For practical purposes, therefore, diagnostic tests can usually be expected to do no more than to cover completely the most important of the detailed facts and items of information contained in a given body of subject matter

and to include a good sampling of each other portion of the content.

The diagnosis of pupils.—The latter part of the preceding section, as well as many discussions to be found elsewhere, may seem to imply that diagnosis refers to the subject matter studied by pupils and their errors therein. This viewpoint, however, is too limited. The problem of diagnosis may be thought of as twofold. One phase is concerned with diagnosis of the pupil himself and the other with that of his knowledge of or ability in a particular body of subject matter. The ordinary classroom teacher, especially in high school where she ordinarily deals with the same pupils in only one or perhaps two classes, is more concerned with the latter, but the former is of considerable importance to her also. Although the principal and other supervisory officials and personnel workers should be expected to take the lead in general diagnosis of the pupil, all teachers should be interested therein both because their help is needed and because the results that may be secured therefrom can be made valuable to them in dealing with pupils in various ways and in diagnosing their achievements. This diagnosis of pupils as contrasted with that of their achievements may be expected to furnish the answer to questions concerning pupils' general attitudes toward school and also toward particular subjects, why they are interested much more in some than in others, or why more in one phase of a subject than in another phase of the same subject, why they are not doing satisfactory school work in general, why they present disciplinary problems, why they leave school, and so on. The effects of such factors as intelligence, emotional traits, study habits, home conditions, health factors, and so forth, will be discovered.

From the standpoint of making thorough diagnoses of this type it is impossible to gather too much information concerning pupils provided it does not become unwieldy by its very bulk. The ages of pupils, their scores upon intelligence and achievement tests, upon tests or rating scales of personality and character, and upon various other sorts of tests, are all useful. Measures of reading ability, both measures, if possible, and teachers' observations of study habits, information as to interests of many sorts, vocational and avocational, scholastic and non-scholastic, and so on, are all frequent sources of insight. Children's health and physical traits

should likewise be studied. Such factors as poor eyesight that has not been recognized and provided for, poor hearing, lack of sufficient sleep, malnutrition, many other physical defects or conditions affecting vitality and health exist in the case of many pupils and have a pronounced effect upon the quality and quantity of their school achievement.

One of the most fruitful sources of information helpful in the diagnosis of pupils is through skillful and careful observation of the pupils, especially when they are unaware that they are being observed. By watching a pupil at study much can frequently be learned about the manner in which he attacks his work. It can be determined, for example, if he has the uneconomical habit of studying for only a moment or two, then spending some time in watching his neighbors, looking out of the window, drawing pictures, or entertaining himself in some other way before continuing his study. Considerable knowledge concerning reading habits, especially about such factors therein as lip movements and confused eye movements, can be gained by observation. In many cases this can profitably be supplemented by talking with pupils concerning their habits of work, by having them in so far as possible "think out loud," that is, endeavor to reproduce aloud for the benefit of the teacher all of their mental processes in dealing with a particular unit of subject matter or exercise.

In connection with the observation of pupils it is frequently helpful to contrast the pupils who appear to be particularly in need of help in certain respects with those who excel in these same respects and then note the differences. For example, the reading habits of the best readers and those of the worst may be contrasted to determine wherein the latter do not follow the same practices as the former. The method of solving problems of the ablest and the poorest pupils in algebra or of attacking a foreign paragraph to be translated of the best and worst in Latin or French may be contrasted. One caution needs to be observed in doing this, however. This is that it does not always follow that because superior pupils use certain methods the same ones are suitable for inferior pupils. The methods that the former use may be suited to them because of their relatively high intelligence or because they have a comparatively good knowledge of the subject matter dealt with and not be appropriate for those who are inferior in intelligence

or in knowledge of subject matter. Indeed, it is frequently better in making contrasts of the sort just referred to to do so for pupils who are as nearly alike as possible in intelligence and perhaps in other qualities, but differ in regard to the thing it is desired to study. Thus, for example, in seeking to help a dull pupil who appears not to know how to memorize satisfactorily, his method of memorization should be contrasted with that of some other dull pupil who appears to have a relatively satisfactory method. To give another example, if an average pupil appears not to have learned how to proceed to translate Latin sentences into English, his method should be contrasted with that of other average pupils who are doing as well as could be expected in this regard.

Sometimes it may not be a question of comparing the worst with the best pupils, but of comparing those at the two extremes with regard to some characteristic or habit of which both extremes are undesirable. In other words, the desired result is the determination of a middle way that avoids the faults of both extreme ones.

The diagnosis of pupils' achievement.—Much of what has been said concerning the diagnosis of pupils themselves applies to the diagnosis of their achievements also. There are, however, a number of additional considerations and facts. In the first place it is impossible to diagnose a pupil's achievement or knowledge of subject matter until he has had some opportunity to learn subject matter of the type to be diagnosed. For example, one cannot diagnose the status of a pupil's achievement in Latin until he has had some opportunity to study Latin, nor his knowledge or ability in physics until he has had opportunity to acquire some such knowledge or ability. It is, however, possible and desirable in many cases to diagnose his status with regard to other subjects more or less contributory to success in the subject in question. For example, pupils beginning French may be diagnosed with regard to their knowledge of English grammar, also of Latin if they have studied it, and those beginning physics with regard to their ability in arithmetic, algebra, and geometry. Most of the diagnosis to be made by teachers, however, should deal with information and abilities supposed to have been acquired during the studies of the high-school subjects themselves rather than of others preparatory to them.

Because of the fact stated in the last sentence coupled with the

additional one that most high-school subjects do not continue throughout several years and do not involve any considerable repetition or review of the same topics, the problem of diagnosing pupils' achievements is more difficult in the high school than in the elementary school. The courses of study pursued in most elementary-school systems provide that in the upper grades, usually the seventh and eighth, there is a more or less complete review of the arithmetic, geography, history, language, physiology, and so forth, previously studied in the lower grades. Thus when a teacher begins a certain topic in any one of these subjects, she can well diagnose the state of her pupils' knowledge and ability therein and use the results as a guide when starting her instruction. In the case of most high-school subjects, however, the same topics do not reappear at more advanced stages of the subject except as they are incidental to others. Because of this, diagnosis in high school must usually be made at more frequent intervals and consist of smaller units of subject matter than in the elementary school. A teacher should plan to spend perhaps 75 or 80 per cent of the time that can be devoted to a given topic upon ordinary or original study and recitation, then diagnose her pupils' achievements therein and use the remainder of the time available for this topic for the remedial measures indicated as needed.

In dealing with achievement much more than with pupils themselves it is possible to do considerable of the work of diagnosis for groups rather than individuals. It very frequently occurs that most if not all of a group of pupils have failed to learn particular points or to acquire certain skills and, therefore, the tests reveal that practically all of the group are weak in certain points. However, it will never, or practically never, occur that group diagnosis is sufficient. It is instead merely a starting point and should be followed up by diagnosis of the individual weaknesses of the pupils in addition to those more or less common to the group. Furthermore, even though the group as a whole needs attention with regard to certain items, it does not follow that all members of the group should receive the same attention. This is merely another way of saying that the procedure in diagnosis and remedial instruction ought to be adapted to the individual's capacities just as all other instruction should be so adapted.

The actual procedure with respect to diagnosis should ordi-

narily be to give a somewhat although not completely general test or, in other words, a test diagnostic with regard to the chief phases or portions of the topic being dealt with. If this reveals that most pupils have acquired reasonably satisfactory knowledge on certain phases, these may for the time being at least receive little or no attention. More detailed diagnostic tests should then be given over the topics in which marked weaknesses or lacks appear. Although a number of standardized diagnostic tests are available in most high-school subjects, the number is not nearly sufficient, so that teachers should supplement them by making many of their own. The tests employed, whether standardized or home-made, should in so far as possible not merely show what errors pupils make or what gaps there are in their knowledge, but why the errors are made or the gaps exist. For example, a diagnostic test dealing with equations in algebra should be so constructed as not merely to show that pupils have difficulty with transposition, for example, but to bring out just what the difficulty is, whether they do not know when to transpose, forget to change signs when transposing, or something else.

Another quality of satisfactory diagnostic tests is that they should frequently contain several questions or exercises dealing with the same facts or processes. The purpose of this is that teachers may know from the results obtained whether or not pupils really know or do not know the points involved. If, for example, a test on equations involves transposition only once, a pupil who transposes correctly may just happen to have done so that time, or one who transposes incorrectly may just happen to have forgotten to change the sign although usually he remembers to do so. If, however, there are several problems involving transposition not so complicated by other procedures that the pupil's responses cannot be analyzed to show his ability in transposing, it will ordinarily be evident whether he knows how to transpose correctly or not. Similarly if a pupil responds correctly to an exercise calling for the proper one of several forms, he may have happened to guess the right answer, but if he responds correctly to several dealing with the same point it is pretty sure that he really knows it.

The next step after securing this detailed information about pupils as well as classes is to take desirable remedial measures. These usually consist of explanations of the points involved fol-

lowed in the case of facts or information by practice or drill exercises adapted to the needs of the various pupils. In other instances exercises calling for application, for reasoning and judgment, and so forth, should be employed. After this there should be another test over the same content to determine what has been the effect of the remedial instruction. In most actual classroom situations there will probably be time for little more than this, that is, one diagnostic test or series of tests of a topic, some remedial instruction and then a follow-up test; sometimes even the follow-up test may not be practicable. It is desirable, however, to go even further and continue the process of alternating diagnostic tests with remedial measures until it is evident that the topics being dealt with have been learned practically as well as can be expected.

To attempt to go into complete detail in connection with diagnosis and remedial instruction would require what would be equivalent to a considerable portion of a book on the teaching of each of the high-school subjects. In the chapters of this book devoted to tests in the various high-school subjects there will be found from time to time descriptions of a number of diagnostic tests and likewise of a few series of practice tests and exercises. Many more of the latter and much other similar material can be purchased from publishing houses and elsewhere, but for some time to come, at least, it will be necessary for teachers to make a considerable portion of their own material. Indeed, it will probably always be desirable for them to do so, since that to be obtained from commercial sources will probably never be well enough adapted to the subject matter taught by each teacher, and to the group of pupils with whom she is dealing, to be entirely satisfactory.

Summary.—The giving of tests in itself is of little value, but should be followed by diagnosis and remedial measures. This is generally true even if tests are given for other purposes also. The tests available for diagnostic purposes vary from those only very slightly diagnostic to those completely so, that cover the smallest items or details of the subject matter concerned. There are, however, comparatively few that reach this degree of perfection; most of those available are only diagnostic to a certain degree. Diagnosis may be thought of as of two types, the diagnosis of pupils and of their achievements. The former is concerned with

diagnosing pupils' general attitudes, habits of study, interests, intelligence, personality, health, and other characteristics that affect the quantity and quality of school work done by them. Such diagnosis should be based upon tests of various sorts, pertinent data concerning pupils and skilled observation of pupils. The diagnosis of pupils' achievement suffers certain disadvantages in high school as compared with elementary school, the chief one being that the same topics do not recur so often and thus the fitting in of diagnosis with the other work is more difficult. The beginning of this type of diagnosis may be group, but for it to be at all satisfactory, much of it must deal with individual pupils. Its purpose is to discover the lacks in their knowledge or ability of the subject in question and if possible why these exist. After this has been determined, remedial measures, ordinarily explanation followed by practice or drill, application, and so on, adapted to the needs shown and also to the individual pupils, should be carried out. If possible a test similar to the first diagnostic test should be given after the remedial instruction to determine its success and whether still further remedial work is needed. Tests suitable for this purpose are not commercially available in sufficient quantities for all phases of all subjects, therefore the teacher should supplement those that can be purchased with many of her own.

BIBLIOGRAPHY

Buckingham, B. R. "Reaching the Individual," *Research for Teachers.* New York: Silver, Burdett and Company, 1926, Chapter IX.

Chapman, J. C. "The Use of Achievement Tests in Diagnosis of Instruction," *Twelfth Annual Schoolmen's Week Proceedings.* Philadelphia: University of Pennsylvania, 1925, p. 245–51.

Greene, H. A. and Jorgensen, A. N. *The Use and Interpretation of Educational Tests.* New York: Longmans, Green and Company, 1929, Chapters VIII, IX, X, and XI.

Levine, A. J. and Marks, Louis. *Testing Intelligence and Achievement.* New York: The Macmillan Company, 1928, p. 214–22.

McCall, W. A. "Measurement in Diagnosis," *How to Measure in Education.* New York: The Macmillan Company, 1922, Chapter III.

Monroe, W. S. *An Introduction to the Theory of Educational Measurements.* Boston: Houghton Mifflin Company, 1923, p. 46–51, 177–79, 245–51.

Monroe, W. S., DeVoss, J. C., and Kelly, F. J. *Educational Tests and*

Measurements, Revised and Enlarged Edition. Boston: Houghton Mifflin Company, 1924, p. 74–89.

Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction*. Yonkers, New York: World Book Company, 1927, p. 18–27, 64–65.

Spencer, P. L. "The Improvement of Teaching by Means of 'Home-Made' Non-Standard Diagnostic Tests and Remedial Instruction," *School Review*, 31:276–81, April, 1923.

Van Wagenen, M. J. *Educational Diagnosis and the Measurement of School Achievement*. New York: The Macmillan Company, 1926. 276 p.

CHAPTER XXIV

STATISTICAL METHODS

Tabulation and classification.—It is the purpose of this chapter to present briefly the statistical methods and procedures most commonly employed in handling test scores and similar data. They will not be dealt with in such complete fashion as would be the case in a complete text on educational statistics and in many cases exceptions and other minor points will be entirely omitted. Most of the terms and measures treated will be those that have already been used and briefly defined in the previous chapters.

The first and in many ways the most important topic that arises in connection with handling such data as test scores is that of tabulation and classification. In dealing with a small class or group of pupils this may not be essential, but practically all teachers have occasion to deal with scores from twenty-five or thirty individuals or more, and also to interpret tabulations of results from still larger numbers of pupils. To approach the subject let us begin with a typical situation. Suppose, for example, that the twenty-two pupils in a French class make the following scores upon a twenty-five word vocabulary test: 16, 19, 23, 14, 17, 18, 20, 19, 16, 13, 24, 21, 17, 25, 16, 14, 20, 22, 14, 9, 23, 20. By examining this series of scores one can in a few moments determine that the highest is 25 and the lowest 9, and in addition gain some idea as to how the rest of the scores run. He will, however, be helped in gaining an understanding of the total situation if the scores are rearranged in order, thus: 25, 24, 23, 23, 22, 21, 20, 20, 20, 19, 19, 18, 17, 17, 16, 16, 16, 14, 14, 14, 13, 9.

Arranging in order need not constitute the final step, however. The series may be condensed somewhat and perhaps made still easier to carry in mind if instead of writing the twenty-two actual scores in order one merely records the different numbers of words correctly answered and the number of pupils getting each.

The series then appears as in Table V. This is known as a grouped or tabulated series or a frequency distribution in contrast to the list of separate scores which is called a simple or ungrouped series. As shown a frequency distribution consists of one column in which the various scores are indicated and another in which the number of individuals receiving each score may be found. This second column is commonly headed "f," since each entry therein is called a frequency, and the total of this column, which, of course, represents the total number of cases or individuals, is denominated "N" for number. The numbers 15, 12, 11, and 10 might be inserted at the proper places in the first or score column with a zero after each in the frequency column,[1] but practice in this regard is not uniform.

TABLE V. TABULATION OF SCORES WITHOUT GROUPING

	f
25–	1
24–	1
23–	2
22–	1
21–	1
20–	3
19–	2
18–	1
17–	2
16–	3
14–	3
13–	1
9–	1
	$N = 22$

The tabulation in Table V, however, is little if any simpler and easier to carry in mind than the complete series of the scores when arranged in order of size. For only twenty-two cases there is little gain in grouping scores unless only a very few groups are used. In this case they might, for example, be grouped as in Table VI. This distribution shows that six of the pupils had scores from 21 to 25 inclusive, eleven from 16 to 20, and so on, and is so short that it is much more easily kept in mind than the complete simple series or the former grouped series. Even here, however, the gain is not very great as the ordinary teacher can comprehend as many as twenty-two scores arranged in order well enough for most practical purposes.

TABLE VI. GROUPED TABULATION OF SAME SCORES AS IN TABLE V

	f
21–25	6
16–20	11
11–15	4
6–10	1
	$N = 22$

If the number of scores or cases is large, the situation becomes different, however. Probably not one person out of a hundred can carry in mind fifty or seventy-five separate scores even if they are arranged in order well enough to retain a very clear picture

[1] The reason this is done is to make clear that there are no scores of these sizes, a fact that might be overlooked unless especially pointed out.

of the whole series. For example, suppose that eighty pupils take a hundred-word test and make the following scores:

94, 89, 72, 61, 77, 84, 88, 58, 93, 79, 71, 68, 82, 64, 81, 90
69, 76, 82, 72, 75, 59, 68, 88, 87, 44, 73, 75, 91, 67, 66, 84
52, 58, 78, 65, 83, 85, 65, 60, 70, 61, 74, 74, 87, 62, 86, 57
92, 80, 73, 77, 78, 63, 83, 55, 72, 84, 68, 89, 83, 73, 78, 66
71, 78, 62, 87, 84, 60, 52, 95, 79, 80, 63, 83, 75, 86, 58, 74

One can look through the series and see that the highest score is 95, the lowest 44, and that most of the scores appear to be between 60 and 90. This, however, is not sufficient to yield a very definite understanding of the situation. If they are arranged in order the task is made somewhat lighter but is still decidedly difficult. The use of a grouped tabulation showing how many persons earned each score made is also not very helpful as forty-two different scores were received by from one to four pupils each and such a series is still too long to be easily comprehended. However, if the scores are grouped by fives as shown by the accompanying Table VII, they are much more easily grasped and also remembered. This shows that one person has a score of from 95 to 99, inclusive, five

TABLE VII. GROUPED TABULATION OF SCORES GIVEN ON PAGE 556

	f
95– *	1
90–	5
85–	10
80–	13
75–	12
70–	12
65–	9
60–	9
55–	6
50–	2
45–	0
40–	1
	$N = 80$

* The reader will note that in this frequency distribution the groups or classes of scores have been indicated merely by giving the lowest score contained in each followed by a dash. This practice is very commonly followed for the sake of economy of space. In such cases it is to be understood that each group of scores includes all from the one given, which is called the lower limit of the class, up to the highest possible score still smaller than the lower limit of the next class. Thus the second group or class, which is labelled 90– and is known as the 90– class, includes all scores from 90 up to and including 94, or, if there are fractional scores, up to and including 94.99. . . . In a distribution so expressed it is always understood that the upper class extends over the same width or distance, that is, includes the same range of scores, as the others. Thus in this case the upper or 95– class is to be understood as extending from 95 up to 99, or, if there are fractions, to 99.99. . . .

have scores of from 90 to 94, ten of from 85 to 89, and so on down.

It is, of course, possible to condense the scores in question still more and group them by tens or twenties, for example. However, in grouping scores one must be guided by two conflicting principles and compromise between what they call for. One has already been suggested, that it is desirable to condense or summarize scores as much as possible to make them more easily comprehended. The other is that too great a degree of their original accuracy and exactness must not be sacrificed by too great summarization. To cite a rather extreme example, suppose that these scores are grouped by twenties. The distribution is then as shown by Table VIII. From these figures one would have no means of knowing that of the twenty-nine scores in the 80– class by far the most are in the lower half of that class, and that only one is at 95 or above, or that of the nine scores in the 40– class all but one are in the upper half. It is generally considered that the most desirable compromise between these two principles is attained when measures are grouped into from ten to twenty classes. Sometimes this is narrowed down to from twelve to fifteen.

TABLE VIII. EXTREMELY SUMMARIZED GROUPED TABULATION OF SAME SCORES AS IN TABLE VII

	f
80–	29
60–	42
40–	9
	$N = 80$

Averages.—In addition to arranging or grouping scores so that the whole series may be fairly easily comprehended, it is helpful in comprehending and describing them to determine some particular score which tends to be representative of the group as a whole. Ordinarily such a score should be an average, using the word in its broad sense to refer not merely to the average or arithmetic mean with which every reader is familiar, but to all measures which indicate central points or scores about which a series of scores or data tend to group themselves.[a]

The mid-score and the median.—The most commonly employed measures for the purpose just indicated in connection with test scores are the mid-score or mid-measure and the median (Md. or Med.). Both may be defined as the point on each side of which half of the scores fall or, in other words, as a point such

[a] In statistical parlance, all such measures are called measures of central tendency.

that half of the scores are above it and half below it. They differ in that the mid-score is such a point for a simple or ungrouped series and the median for a tabulated series.[3]

The mid-score is, of course, very easily found. For example, to find the mid-score of the series of twenty-two scores on the vocabulary test given on page 554, all that one must do is to find the point above and below which are eleven of the twenty-two scores. Taking this series as arranged in order and counting down from the highest end, one finds that the second score of 19 is the eleventh; counting up from the lower end 18 is found to be the eleventh score. Thus any point between 18 and 19 fulfills the condition of having half of the scores above and half below it. The conventional practice in such cases is to take the point halfway between the two scores or, in other words, the average of the two, which is 18½. Thus the mid-score or mid-measure of the given series is 18½. If another score had been added, for example 8, so that the series contained twenty-three scores, the mid-measure would be the twelfth score which, of course, is the same regardless of the end from which one counts. Therefore 18 would be the mid-score. To state the rule more generally, one may find the mid-score of an odd number of scores by arranging them in order and selecting the one on each side of which there are an equal number and of an even number of scores by taking the average of the two mid-most ones.

It is, of course, possible to apply the same method to the series of eighty scores given on page 556 and, having arranged them in order, to determine the mid-score. In this case it is found that there are thirty-eight scores larger than 75 and thirty-nine smaller, with three at 75. Therefore the two mid-most scores, that is, the fortieth score counting in from the high end and also the fortieth from the low end, are both 75 and their average, of course, 75, hence 75 is the mid-score.

In this case, however, the labor of arranging the eighty scores in order is not much if any less than that of tabulating them. Therefore it would be about as easy to determine their median as their mid-score. The method of doing so is in theory the same as that of determining the mid-score, but the details are somewhat

[3] The term "median" is often used to include both the mid-score and the median, but it is better not to do so.

different. Since half of the eighty cases, or forty, must be below the median, one starts at the lower end of the frequency column and counts up, adding all the frequencies until one secures the largest possible number not greater than forty. Thus in this case one adds as follows, starting at the lower end of the frequency column in the distribution on page 556: $1 + 0 = 1$, $1 + 2 = 3$, $3 + 6 = 9$, $9 + 9 = 18$, $18 + 9 = 27$, $27 + 12 = 39$. In other words, all thirty-nine cases in the classes up to and including the 70– class are below the median. Also there is still one more case below it. This, of course, is the lowest case in the 75– class. The exact value of the lowest score in the 75– class is not known, but since there are twelve cases in this class, the best estimate as to the location of the median is that it is one-twelfth of the distance up from the bottom of this class.[4] Since a class covers a range of five score points, one-twelfth of five, which is .42, is added to 75, giving a value of 75.42 for the median. One may arrive at the same result by counting down from the upper end. Doing so, $1 + 5 = 6$, $6 + 10 = 16$, $16 + 13 = 29$, the total number of cases down through the 80– class. Therefore the eleven largest of the twelve scores in the 75– class are above the mean, and eleven-twelfths of the width of that class or eleven-twelfths of five is subtracted from 80, giving 75.42, the same result already obtained, for the median.

Perhaps the computation of the median can be made clearer by a second example, Table IX. There are fifty-four cases in the accompanying tabulation, therefore the median is the point above and below which there are twenty-seven of them. Starting at the bottom $4 + 0 = 4$, $4 + 1 = 5$, $5 + 0 = 5$, $5 + 2 = 7$, $7 + 3 = 10$, $10 + 5 = 15$, the largest number that can be obtained in this manner without exceeding 27. Subtracting 15 from 27, 12 is found to be the number of cases in the next higher, or 70–, class below the median. Since there are sixteen cases in

TABLE IX. COMPUTATION OF THE MEDIAN

	f
90–	6
80–	17
70–	16
60–	5
50–	3
40–	2
30–	0
20–	1
10–	0
0–	4
	$N = 54$

$$\text{Md.} = 70 + 10 = 77.5 - \frac{\frac{54}{2} - 15}{16}$$

[4] It is assumed that all the cases in any given class are equally distributed throughout the distance covered by that class.

that class, twelve-sixteenths of 10, the class width, or interval, must be added to 70, giving 77.5 for the median.

In formula form the method of finding the median may be expressed thus: $\text{Md.} = l + \frac{\frac{N}{2} - S}{f} i$. In this l is the lower limit of the class within which the median falls or, in other words, of the next class above the one with which one stops adding the frequencies. N is, as usual, the total number of cases. S denotes the so-called "partial sum," that is, the sum obtained by adding the frequencies until the largest possible number that is not greater than $\frac{N}{2}$ is secured; f is the frequency in the next class above the last one whose frequency is added in getting S, that is, in the same class of which l is the lower limit; i is the abbreviation for interval or class interval, the width of the class or the number of score points included within the class. Applying this formula to the distribution of eighty scores previously used the median is found to be 75.42, as already shown, thus:

$\text{Md.} = 75 + \frac{\frac{80}{2} - 39}{12} 5 = 75.42$. For the distribution of fifty-four cases the formula gives $\text{Md.} = 70 + \frac{\frac{54}{2} - 15}{16} 10 = 77.5$.

It was suggested above that in determining the median one may either count upward from the bottom or downward from the top of the distribution. Probably because there is slightly less danger of error, it is customary to count from the bottom up, but one may also employ a modification of the formula which provides for counting from the top down. Sometimes it is recommended that this be used as a check on the other. It is as follows:

$\text{Md.} = u - \frac{\frac{N}{2} - S}{f} i$. In this there are no new terms except u, which is used for the upper limit of the class in which the median lies or, in other words, of the class immediately below the last class of which the frequency was used in counting down. This may be used in the two examples already employed as follows:

$$\text{Md.} = 80 - \frac{\frac{80}{2} - 29}{12}\ 5 = 75.42,$$

and

$$\text{Md.} = 75 - \frac{\frac{54}{2} - 23}{16}\ 10 = 77.5.$$

When employing either formula for the median it is most convenient to begin the substitution with $\frac{N}{2}$, then count the frequencies and determine S, after which l or u and f and finally i can be supplied.

In connection with the computation of the median there are several types of more or less exceptional cases which occasionally occur. Only one of them, however, is met with frequently enough that it seems worth while to mention it here, although several others will be found in many of the textbooks on statistics. This case is illustrated by the distribution in Table X. Applying the formula as suggested above, $\frac{N}{2}$ is found to be 35. Next starting to count up the frequencies from the bottom we find that $8 + 6 = 14$, $14 + 4 = 18$, $18 + 5 = 23$, $23 + 3 = 26$, $26 + 4 = 30$, $30 + 5 = 35$, which is exactly the same as one-half the total number of cases. According to the formula, $\text{Md.} = 45 + \frac{\frac{70}{2} - 35}{7} 5 = 45$. In other words, since $S = \frac{N}{2}$, the fraction in the formula becomes zero and the median is the lower limit of the next class. Since this is always true if $S = \frac{N}{2}$, it is unnecessary to go through the form of substituting in the formula when this is discovered. Instead one can immediately know that the lower limit of the next class is the median. To make this

TABLE X. COMPUTATION OF THE MEDIAN WHEN $\frac{N}{2} = S$

	f
55–	16
50–	12
45–	7
40–	5
35–	4
30–	3
25–	5
20–	4
15–	6
10–	8
	$N = 70$

$$\text{Md.} = 45 + \frac{\frac{70}{2} - 35}{7}\ 5 = 45$$

clear another illustration, Table XI, is given. In this $\frac{N}{2} = 32$, and counting the frequencies up from the bottom we have $3 + 4 = 7$, $7 + 4 = 11$, $11 + 8 = 19$, $19 + 13 = 32$, therefore one knows at once that the median is the lower limit of the next class, or 15.

TABLE XI. DETERMINATION OF THE MEDIAN WHEN $\frac{N}{2} = S$

	f
27–	2
24–	4
21–	5
18–	7
15–	14
12–	13
9–	8
6–	4
3–	4
0–	3
	$N = 64$

$S = 32 = \frac{N}{2}$, therefore

Md. = 15

The quartiles.—Although the median or some other average is generally the best single score to choose as representative of the whole group, yet it alone lacks much of giving a complete or satisfactory description thereof. The reason is that it indicates nothing as to how widely the group of scores spreads or scatters about it. Two classes may have equal medians but in one the scores may range over several times as many points as in the other. One way of giving information supplementary to the median which will show something of how much the scores spread out is to indicate certain other fixed points in the distribution. Of these the first and third quartiles (Q_1 and Q_3), also known as the lower and upper quartiles, are the most commonly employed. As the name implies, the first quartile is the point below which are one-fourth of the scores and above which are three-fourths, whereas the third quartile is the point below which are three-fourths and above which are only one-fourth.[8]

The first and third quartiles can be found for either a simple series or a tabulated one and in either case the method is similar to that of determining the mid-score and median, respectively. However, it is not very common to compute the quartile points of a simple series, as such series are usually too short to make it very desirable to do so. When they are obtained, it is common to approximate unless the number of cases is such as to yield them readily. Thus the usual method of getting them for the series of twenty-two scores previously employed is to take the sixth score from the bottom as the middle one of the lower eleven and the

[8] The expression "second quartile" is very rarely employed instead of "median." It is, of course, the point below which there are two-fourths or one-half of the scores.

sixth from the top as the middle one of the upper eleven. Doing this, $Q_1 = 16$ and $Q_3 = 21$.

In finding the quartiles of a grouped distribution the same formula is used as for the median except that for the first quartile $\frac{N}{2}$ is changed to $\frac{N}{4}$, and for the third to $\frac{3N}{4}$.

Therefore, $Q_1 = l + \frac{\frac{N}{4} - S}{f} i$, and $Q_3 = l + \frac{\frac{3N}{4} - S}{f} i$. For the latter not infrequently the method of counting down is employed since it necessitates going through only one-fourth rather than three-fourths of the distribution. The formula is

$$Q_3 = u - \frac{\frac{N}{4} - S}{f} i.$$

The application of these formulae may be illustrated by referring again to the first two distributions used in computing the median. For the first, involving the eighty cases, $Q_1 = 65 + \frac{\frac{80}{4} - 18}{9} 5 = 66.11$, and $Q_3 = 80 + \frac{\frac{3 \cdot 80}{4} - 51}{13} 5 = 83.46$ or, working down from the top, $85 - \frac{\frac{80}{4} - 16}{13} 5 = 83.46$. For the second example, given on page 559, $Q_1 = 60 + \frac{\frac{54}{4} - 10}{5} 10 = 67$, and $Q_3 = 80 + \frac{\frac{3 \cdot 54}{4} - 31}{17} 10$, or $90 - \frac{\frac{54}{4} - 6}{17} 10, = 85.6$.

Percentiles.—In describing or indicating how much or little a distribution of scores spreads out about the median or some other average, it is frequently desirable to have more than two other points indicated. For this purpose it is possible to use any division points one wishes. Occasionally tertiles, that is, points which divide a distribution into three equal parts, are employed; quintiles, which divide it into five parts; sextiles, which divide it into six; deciles, which divide it into ten; and others. However,

by far the most common, after the median and quartiles, are percentiles, that is, points that divide a distribution into one hundred parts, each containing the same number of cases. Generally not all, or even many, of the percentile points are found for a distribution. In the case of test scores seven, the fifth, tenth, twenty-fifth, fiftieth, seventy-fifth,[6] ninetieth, and ninety-fifth, are often determined, thus showing the scores below which 5 per cent, 10 per cent, and so on, respectively, of the pupils fall. Sometimes every tenth percentile is given, with perhaps the fifth and ninety-fifth also.

The method of finding a percentile is the same as that of finding a median or quartile, the only difference being that the first term in the fraction is found by taking the given per cent of N instead of one-half, one-fourth, or three-fourths of it. In generalized form, letting "x" stand for the percentile desired, the formula may be written as follows: $Per_x = l + \frac{\frac{xN}{100} - S}{f} i$. For the fifth percentile, for example, the formula becomes $P_5 = l + \frac{\frac{5N}{100} - S}{f} i$; for the tenth percentile it is $P_{10} = l + \frac{\frac{10N}{100} - S}{f} i$, and so on. To illustrate, the fifth and tenth percentiles of the distribution of eighty scores used a number of times previously are found as follows: $P_5 = 55 + \frac{\frac{5 \cdot 80}{100} - 3}{6} 5 = 55.83$, and $P_{10} = 55 + \frac{\frac{10 \cdot 80}{100} - 3}{6} 5 = 59.17$. To illustrate further the twentieth and ninetieth percentiles for the distribution on page 559 are $P_{20} = 60 + \frac{\frac{20 \cdot 54}{100} - 10}{5} 10 = 61.6$, and $P_{90} = 90 + \frac{\frac{90 \cdot 54}{100} - 48}{6} 10 = 91$.

[6] The twenty-fifth, fiftieth, and seventy-fifth percentiles are, of course, the same as the first quartile, median, and third quartile, respectively.

[7] Both *Per* and *P* are commonly used as abbreviations for percentile. The writer has no especial preference, though perhaps the latter should be used as being shorter.

In this latter case, as indeed in cases where the percentile desired is very much above fifty, it would probably be easier to use the method suggested for the third quartile and also as a check upon the median and work downward from above. For the ninetieth percentile, therefore, one might go down 10 per cent of the distance, thus $P_{90} = 100 - \frac{\frac{10 \cdot 54}{100} - 0}{6} 10 = 91$. To give a second example of this same procedure, the eightieth percentile of the same distribution may be found as follows: $P_{80} = 90 - \frac{\frac{20 \cdot 54}{100} - 6}{17} 10 = 87.2$.

Changing ranks to percentiles and vice versa.—It not infrequently occurs that for some reason or other it is desirable to change pupils' ranks into percentile scores or vice versa. For example, the same pupil may be a member of an English class of twenty-eight pupils, an algebra class of twenty-six, a general science class of twenty, and a manual training class of sixteen. His ranks in the four classes, counting from the bottom up, may be eight, six, nine, and five, respectively, and it may be desired to know how his rank in each compares with that in the others. If there were the same number of pupils in all four classes, a mere statement of his rank would be sufficient, but since this is not the case, stating his respective ranks does not readily make apparent the desired information. One of the easiest ways of doing so is to change all of his ranks into percentile ranks or scores. The formula for so doing is very simple, being as follows: $P = \frac{100(R - \frac{1}{2})}{N}$ The only new term in this formula is R, which is the pupil's rank, counting one as the lowest rank, two as the next lowest, and so on up. Substituting in this formula we find that the pupil's percentile rank or standing in his English class is $\frac{100(8 - \frac{1}{2})}{28}$, which equals 26.8; in his algebra class it is $\frac{100(6 - \frac{1}{2})}{26}$, which gives 21.2; in general science $\frac{100(9 - \frac{1}{2})}{20}$, which equals 42.5; and in manual training $\frac{100(5 - \frac{1}{2})}{16}$, or 28.1. Thus it appears that com-

pared with the other members of the various classes the pupil in question is doing his best work in general science and his poorest in algebra.

Another reason for changing ordinary ranks to percentile ranks is that it is sometimes desired to average a pupil's ranks. This likewise cannot be done directly if there are different numbers in the groups to which he belongs. Using the same illustration as above, however, the average rank of the pupil can be found by adding his four percentile ranks and dividing by four, which gives 29.7 as his average percentile rank in the four classes.

There are occasional, although less frequent, situations in which it is convenient to reverse the process and change a pupil's percentile rank into his absolute rank in a given group. The formula for this is: $R = \frac{N \cdot P}{100} + \frac{1}{2}$. Thus if it is known that a pupil's percentile rank in a class of thirty is 75, his absolute rank is as follows: $\frac{30 \cdot 75}{100} + \frac{1}{2} = 23$. In other words, he is the twenty-third up from the bottom in the group of thirty.

Attention should perhaps be called to the fact that an individual cannot have a percentile rank of 0 or 100. It will be recalled that a percentile rank or score means that the pupil exceeds the given per cent of the pupils concerned. It is readily apparent that a pupil cannot exceed 100 per cent of the pupils in his group since he cannot actually exceed himself. Similarly he cannot have a percentile rank of 0 since this would mean that 100 per cent of the pupils made scores above his and, of course, this could not be true unless his own score were included in the 100 per cent. The highest percentile rank possible in any given group may be found by the formula $P = 100 - \frac{50}{N}$ and the lowest by the formula $P = \frac{50}{N}$.

It is possible to change ordinary or absolute ranks to any one of many other uniform bases of the same sort as percentile ranks, such as quartile ranks, decile ranks, and so forth. However, percentile ranks are the only ones commonly employed for this purpose. The generalized formula for this purpose is: $X = \frac{x(R - \frac{1}{2})}{N}$, and that for finding the ordinary rank from the

tertile, quintile, decile, or other rank: $R = \frac{NX}{x} + \frac{1}{2}$. In these formulae X indicates the desired type of rank such, for example, as quintile or decile, and x indicates the number of divisions in this rank being used, for example, 5 for quintiles, 10 for deciles, 100 for percentiles, and so on.

The mean.—The term "mean," abbreviated M, is ordinarily applied in statistical discussions to what is called the average in elementary-school arithmetic and in daily life. Sometimes the term "arithmetic" is prefixed to either or both, thus, "arithmetic mean" and "arithmetic average." Every reader is undoubtedly familiar with the method of finding the mean in a simple situation, that is, merely to add the scores or other quantities to be averaged and divide by the number. Thus, if one desires to find the mean of the scores made by the class of twenty-two pupils referred to first on page 554, the scores are added, giving a sum of 400, and this divided by 22, yielding a mean of 18.18. This method, however, unless an adding machine is available, becomes rather laborious for large numbers of cases. Also it cannot be applied in its simplest form in the case of a tabulated or group distribution. The essential feature of the procedure in such a case is to secure the sum by multiplying each score by the number of cases at that score and then adding these products. For example, by employing the tabulated distribution of the twenty-two scores[*] one can secure the mean by adding the products of 1 times 25, 1 times 24, 2 times 23, and so on, and thus securing the sum of 400 to be divided by 22.

In the case of a frequency distribution, in which several scores have been grouped together in a single class, the situation is complicated by the fact that one does not know the exact values of the scores but must rather assume a value for those in each group or class. It is the customary and conventional procedure, though with some exceptions, to assume as this value the midpoint of the class or, in other words, the point half way from the lower limit of the class up to the lower limit of the next class. For example, in the tabulation of eighty scores previously used a number of times, the midpoint of the 90– class is taken as 92.5, since that is one-half of the distance from 90 up to 95, that of the 85– class as 87.5, one-

[*] *See* p. 555.

half of the distance from 85 up to 90, and so on. Each midpoint is then multiplied by the corresponding frequency and the sum divided by the total number of cases. If this procedure is applied in the example just referred to, the result shown in Table XII is obtained. It will be seen that a sum of 5970 is secured which, when divided by the number of cases, 80, yields a mean of 74.625.

TABLE XII. COMPUTATION OF THE MEAN BY LONG METHOD

		f		
97.5	×	1	=	97.5
92.5	×	5	=	462.5
87.5	×	10	=	875.0
82.5	×	13	=	1072.5
77.5	×	12	=	930.0
72.5	×	12	=	870.0
67.5	×	9	=	607.5
62.5	×	9	=	562.5
57.5	×	6	=	345.0
52.5	×	2	=	105.0
47.5	×	0	=	00.0
42.5	×	1	=	42.5
N =		80		5970.0

$M = 74.625$

The method just shown is fairly satisfactory, but it is possible to lessen the amount of computation required somewhat. One way of doing this is shown by the computation in Table XIII. It will be noted that two columns have been added to the original frequency distribution and also a pair of horizontal lines enclosing the 75– class drawn through it. The meaning of these will be explained. The first step in employing this method is to estimate or guess the class within which the mean falls. Looking over the distribution one might in this case guess that the mean fell in the 75– class. After such a guess has been made it is customary to draw the parallel lines shown to mark off or distinguish this class.[9] The class in which the mean is assumed to be is known as the zero class. This is shown by *0*, which appears in the third column between the pair of horizontal lines. In the column headed *d*, for deviation or difference, are entered the distances from each class to the zero class. Thus the 95– class, which is

TABLE XIII. COMPUTATION OF THE MEAN BY PARTIALLY SHORTENED METHOD

	f	d	fd
95–	1	+20	20
90–	5	15	75
85–	10	10	100
80–	13	5	65
75–	12	0	+260
70–	12	− 5	− 60
65–	9	10	90
60–	9	15	135
55–	6	20	120
50–	2	25	50
45–	0	30	0
40–	1	35	35
N =	80		−490
		Σfd =	−230

$$M = 77.5 - \frac{230}{80} = 74.625$$

[9] It is not highly important that the guess as to the location of the mean be correct. No error will be involved in the succeeding computations if it is not correct, the only result being to make the figures dealt with slightly larger.

twenty points above the 75– class, has + 20 in the *d* column, the 90– class + 15, and so on. It will be noted that the classes below 75 have negative entries in this column. This is the case because the scores included in them are smaller than those in the 75– class. In practice the easiest way to secure the entries in the *d* column is to start from the zero class and proceed up, entering in the column opposite the next larger class a figure equal to the class interval, above that one equal to twice the class interval, above that one equal to three times the class interval, and so on, as far as the distribution extends. A similar procedure is then carried out downward from the zero place. In the fourth column, which is headed *fd*, that is, the product of the frequency times the deviation, will be found the indicated product for each class. The first entry, 20, is secured from 1 times 20, the next 75, from 5 times 15, and so on. Since all of the entries in this column above [10] the zero class are positive and since the entry for the zero class is always zero, it is customary to sum up those above and write their sum, which is 260 in this case, in the space following the zero. Similarly the sum of all the negative products, which are those below the zero class, is found, in this case being — 490, and then the algebraic sum of the whole column is found by combining the plus sum and the negative sum, yielding in this case a result of — 230. This is frequently labelled Σfd, Σ being a symbol meaning "the summation of." By dividing Σfd by the total number of cases, the average error in guessing the mean may be found. In this case — 230 divided by 80 equals — 2.875. Since the assumed mean was taken in the 75– class and, furthermore, is always at the midpoint of its class,[11] it is 77.5. Subtracting 2.875 from this gives a mean of 74.625, the same as was previously found by the other method.

[10] In speaking of the distribution and accompanying computations, it is assumed that the distribution is arranged as are all those in this book, with the larger scores at the top and the smaller ones at the bottom. If this arrangement is reversed, the positive *d*'s and *fd*'s will, of course, be below the zero class, and the negative ones above, in so far as actual placement on the page is concerned.

[11] The assumed mean is taken at the midpoint of its class because, according to the assumption that the measures in a class are distributed at equal distances through it, their mean is the midpoint of the class.

TABLE XIV. COMPUTATION OF THE MEAN BY SHORT METHOD

	f	d	fd
95–	1	+4	+ 4
90–	5	3	15
85–	10	2	20
80–	13	1	13
75–	12	0	+ 52
70–	12	−1	− 12
65–	9	2	18
60–	9	3	27
55–	6	4	24
50–	2	5	10
45–	0	6	0
40–	1	7	7
	$N = 80$		− 98
		80	− 46
		$c =$	−.575
		$i =$	5.
		$ci =$	−2.875
		$Ass.M. =$	77.5
		$M =$	74.625

Although the method just indicated reduces the sizes of the numbers dealt with below those used in the previous method it is possible to go still a step further in this direction. How to do this is shown by the computation in Table XIV. The difference between this and the one in Table XIII is that instead of expressing the deviations or entries in the d column in terms of the actual number of score points, they are taken in terms of classes. In other words, since the 75– class is the zero class, the class immediately above it, the 80– class, has a d value of + 1, the 85– class of + 2, and so on; also the class immediately below it, the 70– class, has a d value of −1, the 65– class of −2, and so on. The fd column is found as before by multiplying each entry in the f column by the corresponding one in the d column. Proceeding as before, summing the plus and minus fd's and taking their algebraic sum, −46 is obtained. Dividing this by 80 gives a result of −.575, the error or correction in the assumed mean. This, however, is expressed not in actual points but in class intervals, and, since the width of each class interval is 5, must be multiplied by 5 to give the error in actual points. Making this multiplication — 2.875 is again secured and, of course, the same mean, 74.625. In formula form this method is expressed as follows: $M = Ass.M + \frac{\Sigma fd}{N} i$, or sometimes $= Ass.M + ci$. In this formula "$Ass.M$" is the symbol for assumed or guessed mean. The "c" in the second one stands for correction, or, in other words, for the quantity obtained when the algebraic sum of the fd column is divided by the total number of cases, and i, of course, is the class interval as previously used.

The use of this so-called "short method" of computing the mean is further illustrated for the distribution originally given on page 559. In this example the assumed mean is taken in the

TABLE XV. COMPUTATION OF THE MEAN BY SHORT METHOD

	f	d	fd
90–	6	+2	+12
80–	17	1	17
70–	16	0	+29
60–	5	−1	− 5
50–	3	2	6
40–	2	3	6
30–	0	4	0
20–	1	5	5
10–	0	6	0
0–	4	7	28
$N = 54$			− 50
			$\Sigma fd = -21$

$$c = \frac{-21}{54} = -.389$$

$$ci = 10 \times -.389 = -3.89$$

$Ass.M = 75.00$

$ci = -3.89$

$M = 71.11$

70– class, or at 75. The sum of the plus fd's is found to be 29, that of the minus ones — 50, giving an algebraic sum or Σfd of — 21. Dividing by N, 54, the correction is found to be — .389 and multiplying by the class interval, 10, $ci = -3.89$. Therefore the mean is 75.00 — 3.89 or 71.11.

The measurement of variability.—Although the degree to which the scores or other measures in a distribution vary from the average may be described by stating other points such as the quartiles, percentiles, and so forth, this is not always a satisfactory way of describing variability. It is often desirable to be able to sum up the variability of a distribution by a single expression rather than to attempt to carry several different points of the distribution in mind. In the following paragraphs the several measures of variability most commonly employed will be explained.

The range.—The range of a distribution of scores is merely the distance from the lowest score in the series up to the highest. It is, therefore, very easily determined if one knows these two scores. If the distribution has already been tabulated so that one does not know the exact highest and lowest scores, the range can nevertheless be determined approximately. However, despite its ease of computation it receives little use because it is highly unreliable as a measure of the spread of the whole group. For example, in the frequently used case of the twenty-two pupils taking the twenty-five word test, the highest score is 25 and the lowest 9, therefore the range is 16. However, if the pupil making a score of 9 had been absent on the day the test was given, the range would have dropped from 16 to 12, since that is the difference between the next lowest score, 13, and 25. In other words, although the whole group would be changed merely by one pupil, the size of the range would be decreased by one-fourth. No measure which depends absolutely on the size or location of only two measures

out of the whole group can be considered very reliable as indicative of the trend of the group.

The quartile deviation.—One of the commonest and most easily computed measures of variability is the quartile deviation (Q), which is sometimes also known as the semi-inter-quartile range. As the second expression indicates, it is half the distance from the first up to the third quartile. In formula form, therefore, $Q = \frac{Q_3 - Q_1}{2}$. Since there are one-fourth or 25 per cent of all the cases between each quartile and the median, the quartile deviation is the average distance from the median that includes one-fourth of the cases in each direction, or one-half in all. Thus, for the distribution of eighty pupils on page 556, $Q = \frac{83.46 - 66.11}{2}$ $= 8.68$. In other words, within a distance of 8.68 from the median, which is 75.42, approximately half of the cases, in this instance forty, will be found. If the distribution were symmetrical or if the distance from the first quartile to the median were the same as that from the median to the third quartile, exactly half would lie within 8.68 of 75.42. Unless a distribution is decidedly unusual in its shape, the per cent of cases included within a distance of Q on both sides of the median is so near fifty that it is commonly assumed to be exactly that.

The interpretation of the quartile deviation may also be extended further. Not only do we know that 50 per cent of the cases in a normal distribution are within $1Q$ of the median and 50 per cent are not, but also that over 32 per cent more, or 82 + per cent in all, are within $2Q$ of the median and 18 — per cent are not, that almost 96 per cent are within $3Q$ of the median and over 4 per cent are not, and so on. In other words, in the example used 82 + per cent of all the cases are within 17.36 points of 75.42, 96 — per cent are within 26.04 points of 75.42, and so on.

For the distribution given on page 559, $Q_1 = 67.0$ and $Q_3 = 85.6$. Therefore $Q = \frac{85.6 - 67.0}{2} = 9.3$, which means that half of the measures are within 9.3 of the median, which is 77.5, or between 68.2 and 86.8. Also, as in the example just above, over 82 per cent of the cases are within 18.6 of the median, or between

58.9 and 96.1, almost 96 per cent within 27.9 of it, or between 49.6 and 105.4, and so on.

Although the quartile deviation does not possess certain statistical refinements and advantages to be found in other measures of variability, it is generally recommended for use in connection with the median and is accurate enough for most practical purposes for which test data are employed by the teacher. Indeed, it is the only measure of variability commonly employed in connection with the median. Also its use is limited in that it is not used along with other averages than the median.

The standard deviation.—Another very common measure of variability is the standard deviation, abbreviated S.D. or, more commonly, σ (sigma). It is computed from the mean and employed in connection with that measure. Slightly over one-third of the cases in a normal distribution lie between the mean and a point one standard deviation from it. In other words, somewhat over two-thirds (68.27 per cent, to be more exact) of all the cases in a normal distribution differ from the mean by no more than the standard deviation. Although these figures do not hold exactly for distributions that are not normal, they are ordinarily assumed to do so and, therefore, are commonly employed in connection with all frequency distributions.

In formula form, for a simple series, $\sigma = \sqrt{\frac{\Sigma d^2}{N}}$.[12] Expressing this in words, the standard deviation is the square root of the average of the squares of the deviations of all the scores in the series from the mean. This definition may sound somewhat difficult, but the actual process of computation is fairly easy. For a simple series, such as the twenty-two scores already often referred to, the method is shown in Table XVI. The first column contains the actual scores. They are summed and the result, 400, divided by 22, the number of scores, yielding a mean of 18.2 if only one decimal is carried. In the second or d column are the deviations or differences of the scores from their mean, 18.2. The third column, headed d^2, contains the squares of the deviations. For example,

[12] This is frequently written $\sigma = \sqrt{\frac{\Sigma x^2}{N}}$, x instead of d being used for a deviation from the mean.

TABLE XVI. COMPUTATION OF STANDARD DEVIATION OF UNGROUPED SCORES BY LONG METHOD

16	−2.2	4.84
19	+ .8	.64
23	+4.8	23.04
14	−4.2	17.64
17	−1.2	1.44
18	− .2	.04
20	+1.8	3.24
19	+ .8	.64
16	−2.2	4.84
13	−5.2	27.04
24	+5.8	33.64
21	+2.8	7.84
17	−1.2	1.44
25	+6.8	46.24
16	−2.2	4.84
14	−4.2	17.64
20	+1.8	3.24
22	+3.8	14.44
14	−4.2	17.64
9	−9.2	84.64
23	+4.8	23.04
20	+1.8	3.24
22 | 400		22 | 341.28
18.2		15.5127

$\sigma = \sqrt{15.5127} = 3.94$

TABLE XVII. COMPUTATION OF STANDARD DEVIATION OF UNGROUPED SCORES BY USE OF ASSUMED MEAN

	d	d^2
16	−2	4
19	+1	1
23	+5	25
14	−4	16
17	−1	1
18	0	0
20	+2	4
19	+1	1
16	−2	4
13	−5	25
24	+6	36
21	+3	9
17	−1	1
25	+7	49
16	−2	4
14	−4	16
20	+2	4
22	+4	16
14	−4	16
9	−9	81
23	+5	25
20	+2	4
22 | 400		22 | 342
18.2		15.5455
		.04
		15.5055

$\sigma = \sqrt{15.5055} = 3.94$

the first score, 16, has a deviation of 2.2 from 18.2, the mean, and the square of 2.2 is 4.84. The sum of the squares is 341.28. Dividing this by 22, their average is found to be 15.5127. The square root of this is 3.94, which is the standard deviation.

Even though the procedure just illustrated is not unduly laborious, it does involve many decimal fractions. By employing an assumed mean, as was done in finding the mean on pages 568–571, these decimals can be eliminated.[12] The application of this formula is shown in Table XVII, the same scores being used. Instead, however, of taking the differences between the scores and the ac-

[12] The formula for this method is $\sigma = \sqrt{\frac{\Sigma d^2}{N} - c^2}$.

tual mean, they are taken from an assumed mean, in this case 18.[14] They are then squared as before, the sum of their squares found and then the average of the squares. In this case it is 15.5455. From this the square of the correction, that is, the difference between the true mean and the assumed mean, is subtracted. This difference is .2 (the difference between 18.2 and 18) and its square .04. Taking this from 15.5455 leaves 15.5055, and the square root of this is 3.94, the same as was found by the first and more laborious method.

To compute the standard deviation of a grouped distribution the latter of the two methods just illustrated is employed. It must, of course, be modified to suit a grouped instead of a simple series in much the same way as is done for the mean. Its computation involves the use of the same form as was employed for getting the mean on pages 571–572, with the addition of one more column and is according to the formula $\sigma = \left(\sqrt{\frac{\Sigma f d^2}{N} - c^2}\right) i$. To illustrate it most of Table XIV is reproduced, with the additional or fd^2 column.

TABLE XVIII. COMPUTATION OF STANDARD DEVIATION OF GROUPED SERIES

	f	d	fd	fd^2
95–	1	+4	+ 4	16
90–	5	3	15	45
85–	10	2	20	40
80–	13	1	13	13
75–	12	0	+ 52	0
70–	12	−1	− 12	12
65–	9	2	18	36
60–	9	3	27	81
55–	6	4	24	96
50–	2	5	10	50
45–	0	6	0	0
40–	1	7	7	49
$N = 80$			− 98	80 | 438
		80 | − 46		$s^2 = 5.4750$
		$c = -.575$		$c^2 = .3306$
				$\sigma^2 = 5.1444$
				$\sigma = 2.27$
				$i = 5$
				$\sigma = 11.35$

[14] If, as here, the mean is known, the assumed mean is commonly taken as the nearest whole number, or perhaps even as the nearest round number.

In this are the squares of the deviations, each multiplied by its corresponding frequency. In actual practice this is most easily done by multiplying each entry in the d column by the corresponding one in the fd column, thus securing d times fd, or fd^2. For example, in the first row of the table, $d = 4$ and fd also $= 4$, so $fd^2 = 16$; in the second $d = 3$ and $fd = 15$, so $fd^2 = 45$, and so on. The sum of this column, 438, is divided by 80, the number of cases. The result, in this case, 5.475, is commonly denominated S^2. From this c^2, here .3306, is subtracted, leaving σ^2, that is, the square of the standard deviation. The square root of this is 2.27, the standard deviation in terms of class intervals. Since the interval is 5, 2.27 is multiplied by that number and the result, 11.35 is the standard deviation in terms of the actual scale or scoring system used.

The computation of the standard deviation may be further illustrated for the distribution of fifty-four cases already employed on page 559 and elsewhere. As in the last example, the

TABLE XIX. COMPUTATION OF STANDARD DEVIATION OF GROUPED SERIES

	f	d	fd	fd^2
90–	6	+2	+12	24
80–	17	1	17	17
70–	16	0	+29	0
60–	5	−1	− 5	5
50–	3	2	6	12
40–	2	3	6	18
30–	0	4	0	0
20–	1	5	5	25
10–	0	6	0	0
0–	4	7	28	196
	54		−50	54 | 297
			54 | −21	$S^2 =$ 5.5000
			$c =$ −.389	$c^2 =$ .1513
				$\sigma^2 =$ 5.3487
				$\sigma =$ 2.313
				$i =$ 10.
				$\sigma =$ 23.13

computation leading to the mean is reproduced, with the fd^2 column in addition. The sum of this last column, 297, is divided by N, which is 54, giving 5.5000 for S^2. From this c^2 is subtracted, leaving $\sigma^2 = 5.3487$. The square root of this, 2.313, is σ, but is in terms of class intervals, so must be multiplied by 10, the interval,

to give the value of the standard deviation in terms of the scale used. This value is, of course, 23.13.

The interpretation of the standard deviation, as well as that of the quartile deviation, may be extended. Not only does a distance of 1σ from the mean include 68 + per cent of all the scores, but also one of 2σ includes 95 + per cent, one of 3σ 99.7 per cent, and so on.

The median deviation.—The median deviation, abbreviated Md.D., is just what its name implies, the median of the deviations. It is not ordinarily determined as one would suspect from this fact, however. Practically never are deviations tabulated and their median found. Instead it is secured by first computing the standard deviation and then multiplying it by .6745.[16] In formula form, $\text{Md.D.} = .6745\sigma$. This relationship between the median and the standard deviation, and hence the formula, is correct only for normal distributions, but is commonly assumed to hold elsewhere and therefore is generally employed.

Applying this formula, the median deviation of the twenty-two vocabulary scores is $.6745 \times 3.94$, the standard deviation, or 2.66. For the series of eighty scores $\text{Md.D.} = .6745 \times 11.35 = 7.66$.

The median deviation is employed in connection with the mean and is interpreted in just the same way as the quartile deviation. Therefore, since the mean of the eighty scores is 74.63, 50 per cent of them may be expected to be within 7.66 of that point, or between 66.97 and 82.29, 82 per cent between 59.31 and 89.95, 96 per cent between 51.65 and 97.61, and so on.

The name "probable error," abbreviated P.E., is very frequently but erroneously used instead of median deviation. This usage is justified when, as sometimes, the deviations concerned are errors, but not otherwise. It is, however, so general and apparently so well established that it is doubtful if it will be dropped. When it is really a measure of error it is customary to write it after the measure to which it applies with a plus-or-minus sign ($\pm$) connecting the two. Thus $M = 48.4 \pm 1.8$, for example, is the same as stating that the mean is 48.4 and its probable error 1.8. In this

[16] This decimal is not exact, but near enough for practical purposes. Indeed, in rough work it is sometimes taken merely as 2/3.

connection a probable error can, if desired, be found for the mean, median, quartiles, standard deviation, or almost any statistical measure.

The coefficient of variability.—The quartile, standard and median deviations are all absolute measures rather than relative ones and therefore do not afford satisfactory means of comparing the variability of different distributions. For example, if one distribution has a median deviation of ten and another of five, the former is twice as variable as the latter on an absolute basis. If, however, the mean of the first is eighty and that of the second twenty, the deviation of the first is only one-eighth of its mean whereas that of the second is one-fourth, so that the second has twice as great proportional or relative variability.

The generally accepted measure of variability is the coefficient of variability or of variation, abbreviated *C. of V.* or just *V.* For reasons that will not be presented here, it is not an entirely satisfactory measure, but no better one has been suggested. The formula for it is simply $100\frac{\sigma}{M}$, the factor 100 being introduced to give a result that is greater than unity. Thus for the series of twenty-two vocabulary scores $C.\ of\ V. = 100\frac{3.94}{18.2} = 21.65$ and for that of the eighty scores it is $100\frac{11.35}{74.63} = 15.21$. These results show that the first is relatively more variable than the second.

Rank scores.—It is often desirable to assign pupils rank scores, that is, to rank them in order from highest to lowest, or vice versa. Although this is a rather common procedure entirely apart from any connection with test scores, it seems desirable to explain briefly just how it should be done. The explanation will be illustrated by using the series of twenty-two vocabulary scores. These are given, in order, in the first of the two columns in Table XX. In the second are the ranks. For most purposes it makes little or no difference whether rank 1 is assigned to the highest or lowest score, but for the sake of consistency with percentile and other similar rankings it is probably best to let 1 be the lowest. In this case, therefore, the score of 9 is given rank 1 and the next lowest, 13, rank 2. Next are three scores of the same size, 14. To determine their rank, we find the average of the next three ranks

TABLE XX. ASSIGNMENT OF RANKS TO SCORES

Score	*Rank*
25	22
24	21
23	19.5
23	19.5
22	18
21	17
20	15
20	15
20	15
19	12.5
19	12.5
18	11
17	9.5
17	9.5
16	7
16	7
16	7
14	4
14	4
14	4
13	2
9	1

after 2. These are 3, 4 and 5, and their average 4, so this rank is assigned to each score of 14. Similarly each of the three scores of 16 receives a rank of 7, the average of 6, 7 and 8. Next are two of 17, which receive ranks of 9.5, the average of 9 and 10. The one score of 18 is rank 11, the two of 19 rank 12.5, and so on to the highest, 25, which is rank 22.

Correlation.—When two series of scores of the same individuals are to be compared, correlation is the method ordinarily used. For example, if the same pupils have taken two algebra tests, or perhaps two forms of the same test, their scores may be correlated to determine how well those made at one testing agree with those at the other. In similar fashion algebra test scores may be correlated with school marks, intelligence quotients, geometry test scores or any other series of measures of the same pupils. There are several more or less common measures of correlation. Most of them range from a maximum value of + 1.00 down through zero to — 1.00. If the correlation or agreement is perfect, that is, if each pupil's score in one series corresponds exactly with his score in the other on the basis of size, the value of + 1.00 is obtained. If the disagreement is as great as possible, the value is — 1.00. If there is no relationship at all, neither agreement or disagreement, it is .00. Intermediate values indicate various degrees of agreement or disagreement according to how nearly they approach + 1.00 or — 1.00.

Although there are a number of different measures of correlation, only the two or three most commonly employed ones will be treated here. These are product-moment correlation and the two varieties of rank correlation.

Product-moment correlation.—This method of computing correlation is by far the most frequently used and may be considered the standard method, although in some situations its value is decidedly limited. Its chief limitation is that it measures ade-

quately only rectilinear or straight-line relationship, that is, relationship expressed satisfactorily by a single degree algebraic equation. To illustrate this the short series of numbers in Table XXI may be employed. In each series are five pairs of numbers, the first of each pair being labelled X and the second Y, for convenience. In the case of each series there is a definite and readily apparent mathematical relationship between the paired numbers in the two columns. In Series A each number in the Y column is just twice as great as the corresponding one in the X column. Expressed algebraically in equation form, $Y = 2X$. In Series B, each Y number is one larger than the corresponding X number, or $Y = X + 1$. Both of these are rectilinear or first degree equations, since they contain only the first powers of X and Y. In Series C, however, each Y value is the square of the paired X value, that is, $Y = X^2$. This is not a rectilinear equation. If the product-moment correlation for each of the three series were found, it would be + 1.00 for both A and B, but somewhat less for C, although there is as close and unvarying a relationship in the latter case as in A and B.

TABLE XXI. PAIRED SERIES ILLUSTRATING PERFECT CORRELATION

Series A		*Series B*		*Series C*	
X	Y	X	Y	X	Y
5	10	5	6	5	25
4	8	4	5	4	16
3	6	3	4	3	9
2	4	2	3	2	4
1	2	1	2	1	1

Although in a general sense any measure of correlation may properly be called a coefficient, the expression "coefficient of correlation," abbreviated r, is conventionally limited to the product-moment measure. It ranges in value as suggested above, from + 1.00 down through zero to — 1.00. A number of methods of calculating it have been devised and are employed, but except for the difference in dealing with simple series and grouped series, the variations are relatively minor and even there it is a question merely of the most convenient form and not of the mathematical operations actually performed. Therefore no attempt will be made to present or explain the various methods of computation, but instead just two will be given, one for simple series and one for grouped series.[16] Both are based upon the general formula

[16] The methods to be explained here are often known as Ayres', although he did not originate them. He did, however, much to bring them to the attention of those who might use them.

$r = \frac{\Sigma xy}{N\sigma_x\sigma_y}$ In this x and y represent the differences or deviations of the scores in the two series concerned from their respective means, so that the numerator is the sum of the products of each corresponding pair of deviations. As shown, the denominator is merely the number of cases times the product of the standard deviations of the two series.

The computation of the coefficient of correlation between two simple series is shown by Table XXII. In the first column thereof

TABLE XXII. COMPUTATION OF COEFFICIENT OF CORRELATION OF UNGROUPED SERIES

X	Y	X^2	Y^2	XY
25	25	625	625	625
24	22	576	484	528
23	24	529	576	552
23	22	529	484	506
22	23	484	529	506
21	22	441	484	462
20	21	400	441	420
20	21	400	441	420
20	15	400	225	300
19	22	361	484	418
19	20	361	400	380
18	17	324	289	306
17	22	289	484	374
17	19	289	361	323
16	19	256	361	304
16	18	256	324	288
16	13	256	169	208
14	17	196	289	238
14	16	196	256	224
14	14	196	196	196
13	16	169	256	208
9	12	81	144	108
400	420	7614	8302	7894

$$r = \frac{7894 - \frac{400 \times 420}{22}}{\sqrt{(7614 - \frac{400^2}{22})\,(8302 - \frac{420^2}{22})}} = \frac{257.64}{\sqrt{341.27 \times 283.82}} =$$

$$\frac{257.64}{311.22} = .83-$$

the same twenty-two scores so often employed previously appear, arranged in order, which is not necessary but merely convenient. In the second column are the scores made by the same pupils on

a second test of the same sort and length. Thus the pupil who made a score of 25 on the first test did likewise on the second, the one who made 24 on the first made only 22 on the second, and so on. According to conventional practice, the first column is headed X and the second Y. To secure the coefficient of correlation, three additional columns are needed. The one headed X^2 has in it the squares of the numbers in the X-column; the one headed Y^2, those of the Y numbers; the last or XY-column contains the products of each X entry and the corresponding Y one. Each of the five columns is totalled and from these sums r is obtained by the following formula, which is merely a variation of the general one already given:

$$r = \frac{\Sigma XY - \frac{\Sigma X \cdot \Sigma Y}{N}}{\sqrt{\left(\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right)\left(\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right)}}.$$

As will be seen, the numerator of the fraction is the sum of the last column minus the product of the sums of the first and second divided by the number of cases. The denominator is the square root of the product of two quantities, of which the first is the sum of the third column minus the square of the sum of the first divided by the number of cases and the second the sum of the fourth column minus the square of the sum of the second divided by the number of this. For the example given

$$r = \frac{7894 - \frac{400 \times 420}{22}}{\sqrt{\left(7614 - \frac{400^2}{22}\right)\left(8302 - \frac{420^2}{22}\right)}} = \frac{257.64}{\sqrt{341.27 \times 283.82}} = \frac{257.64}{311.22} = .83 - .$$

The method just illustrated is as simple and direct as any, but it is possible to lessen the necessary labor somewhat by reducing the sizes of the numbers dealt with. To do so an additional step, but a very easy one, is necessary. This is merely to subtract from all the numbers in each series some number, the same for each series but not necessarily so for the two series. Usually the number subtracted is equal to the smallest score in each series, or, if

one desires to avoid zeros, is one less than that. Sometimes, however, a round number almost as large as the smallest score is subtracted and sometimes even one larger, perhaps about equal to the mean, thus leaving many negative numbers. In this case the smallest scores have been subtracted, 9 for the X column and 12 for the Y one. The reduced numbers so obtained are given in the first two columns of Table XXIII, headed x and y.[17] Proceeding

TABLE XXIII. COMPUTATION OF COEFFICIENT OF CORRELATION OF UNGROUPED SERIES WITH REDUCED SCORES

x	y	x^2	y^2	xy
16	13	256	169	208
15	10	225	100	150
14	12	196	144	168
14	10	196	100	140
13	11	169	121	143
12	10	144	100	120
11	9	121	81	99
11	9	121	81	99
11	3	121	9	33
10	10	100	100	100
10	8	100	64	80
9	5	81	25	45
8	10	64	100	80
8	7	64	49	56
7	7	49	49	49
7	6	49	36	42
7	1	49	1	7
5	5	25	25	25
5	4	25	16	20
5	2	25	4	10
4	4	16	16	16
0	0	0	0	0
202	156	2196	1390	1690

$$r = \frac{1690 - \frac{202 \times 156}{22}}{\sqrt{(2196 - \frac{202^2}{22})\ (1390 - \frac{156^2}{22})}} = \frac{257.64}{\sqrt{341.27 \times 283.82}} =$$

$$\frac{257.64}{311.22} = .83-$$

as before, by squaring the x and y entries and getting their products and substituting in the formula, we have

[17] When X and Y are used for scores, x and y are usually employed to denote their deviations from the mean, assumed mean or other similar point.

$$r = \frac{1690 - \frac{202 \times 156}{22}}{\sqrt{(2196 - \frac{202^2}{22})\ (1390 - \frac{156^2}{22})}} =$$

$$\frac{257.64}{\sqrt{341.27 \times 283.82}} = \frac{257.64}{311.22} = .83 -,$$ just the same as before.

The writer recommends that if the smallest score in a column is as great as one-fourth of the largest score, a subtraction be made as just illustrated, otherwise not. Also he advises that if, as is not often possible, all the numbers in a column or series can be divided evenly by any number greater than one, this be done, the highest common divisor being employed.

The method of computing the coefficient of correlation between two grouped series is the same in principle and theory, but decidedly different in form from that for simple series. To illustrate it, the series of eighty scores first given on page 556 and a similar second series will be used. The two scores of each of the eighty pupils are as follows, those of each pupil being separated by a dash and those of one pupil from those of another by a comma:

94–96, 89–87, 72–69, 61–66, 77–75, 84–83, 88–91, 58–61, 93–94, 79–76,
71–72, 68–65, 82–76, 64–57, 81–85, 90–89, 69–69, 76–77, 82–83, 72–78,
75–74, 59–62, 68–68, 88–90, 87–86, 44–47, 73–77, 75–74, 91–93, 67–71,
66–68, 84–82, 52–53, 58–59, 78–79, 65–64, 83–86, 85–82, 65–64, 60–59,
70–73, 61–62, 74–76, 74–73, 87–91, 62–59, 86–84, 57–62, 92–92, 80–87,
73–84, 77–68, 78–80, 63–62, 83–87, 55–54, 72–76, 84–87, 68–71, 89–93,
83–80, 73–71, 78–80, 66–63, 71–75, 78–80, 62–58, 87–84, 84–87, 60–63,
52–49, 95–94, 79–81, 80–83, 63–66, 83–79, 75–82, 86–90, 58–60, 74–72.

The correlation of these scores could, of course, be found by the method just illustrated, but in view of the number of cases to be dealt with it is better to employ another.[18]

The first step is to enter the cases in a correlation table, also sometimes called a table of double entry. For the scores just given

[18] It is extremely unusual for the simple series method of computing correlation to be employed when there are more than forty or fifty cases concerned.

it seems best to use intervals of five both ways in the table,[19] beginning with the 40– class for the first scores and the 45– class for the second ones. It is customary to consider the first scores as the X scores and the second ones as the Y scores. The coördinate axis system of mathematics is employed, the X scores being laid off horizontally and increasing from left to right and the Y scores vertically, increasing from the bottom up. The blank table wherein to tabulate the scores is, therefore, shown as Table XXIV.

TABLE XXIV. BLANK CORRELATION TABLE FOR SCORES ON PAGE 585

	40–	45–	50–	55–	60–	65–	70–	75–	80–	85–	90–	95–
95–												
90–												
85–												
80–												
75–												
70–												
65–												
60–												
55–												
50–												
45–												

After the construction of the table, the pairs of scores are entered therein. The ordinary way of doing this is to employ a tally mark (|) for each pair, placing it in the appropriate column and row. Thus for the first pair, 94–96, a tally mark should be placed in the rectangle formed by the intersection of the 90– column and the 95– row; for the next, 89–87, it should be in the 85– row and the 85– column; for the third, 72–69, in the 70– column and the 65– row; and so on until all eighty have been entered. The actual tally marks are not shown here, but instead, in the next table, Number XXV, are the figures which show how many should be in

[19] The determination of the intervals to be used is merely a matter of considering each series separately and choosing the interval therefor as suggested on page 557.

TABLE XXV. COMPUTATION OF THE COEFFICIENT OF CORRELATION OF GROUPED SERIES BY MEANS OF CORRELATION TABLE

	40–	45–	50–	55–	60–	65–	70–	75–	80–	85–	90–	95–	T_y	d_y	fd_y	fd^2_y	Σx	Σxy
95–											1		1	11	11	121	11	121
90–										5	3	1	9	10	90	900	95	950
85–									6	2	1		9	9	81	729	85	765
80–							1	5	5	3			14	8	112	896	122	976
75–							5	4	2				11	7	77	539	85	595
70–						3	5	2					9	6	54	324	68	378
65–					2	4	1	1					8	5	40	200	49	245
60–				4	3	3							10	4	40	160	49	196
55–				1	4								5	3	15	45	24	72
50–			1	1									2	2	4	8	7	14
45–	1		1										2	1	2	2	4	4
T_x	1	0	2	6	9	9	12	12	18	10	5	1	80		526	3924	594	4816
d_x	1	2	3	4	5	6	7	8	9	10	11	12						
fd_x	1	0	6	24	45	54	84	96	117	100	55	12	594					
fd^2_x	1	0	18	96	225	324	588	768	1053	1000	605	144	4822					
Σy	1	0	3	21	34	44	78	85	108	92	50	10	526					
Σxy	1	0	9	84	170	264	546	680	972	920	550	120	4816					

$$r = \frac{4816 - \frac{594 \times 526}{80}}{\sqrt{\left(4822 - \frac{594^2}{80}\right)\left(3924 - \frac{526^2}{80}\right)}} =$$

$$\frac{410.45}{\sqrt{411.55 \times 465.55}} = \frac{410.45}{437.72} = .94-$$

each rectangle.[20] The "1" in the 95– column and the 90– row, for example, shows there was just one pupil whose score on the first test was from 90 to 94, inclusive, and whose score on the second was from 95 to 99.

After the entries have all been made in the body of the table, the next step is to total each row and column. The totals of the rows are given in the column headed T_y and of the columns in the row headed T_x.[21] Each entry in this column and row is a frequency, but the heading T rather than f is often used in correlation tables. In the next column, headed d_y, are numbers or deviations beginning with 1 for the lowest class and increasing by one for each class.[22] The fd_y column contains the products of the entries in the T_y column with the corresponding ones under d_y and the fd_y^2 column the products of the entries under d_y and fd_y, just as do the fd and fd^2 columns on page 576. Similarly at the bottom of the table the x-deviations in the d_x row begin with 1 for the 40– class, the lowest one, and increase to the right. The entries in the fd_x and fd_x^2 rows are obtained in the same way as those in the fd_y and fd_y^2 columns. Up to this point nothing not included in the computation of the mean has been done, the only differences being in the use of deviations running up from 1 instead of both ways from zero and in the fact that two distributions instead of one are being dealt with.

The next steps are new, however. For a row at a time, each entry in the table is multiplied by the d_x value of its column, that is, immediately below it, and the sum entered in the column headed Σx. Thus in the first row the only entry, 1, has a d_x value of 11, so Σx for that row is 11. In the second or 90– row there are 5 cases with an x deviation of 10, 3 with one of 11 and 1 with one of 12,

[20] In the actual filling in of correlation tables it is frequently convenient to enter the tally marks with a moderately soft pencil, then write in the figures showing the numbers of marks with ink and finally erase the marks themselves.

[21] The subscript $_y$ is commonly employed, as shown, to indicate the totals, deviations and so forth in columns, that is, in a vertical or y-direction, and $_x$ for those in rows or in an x-direction.

[22] These numbers may begin with zero at the bottom instead of 1, or they may be taken as in the short method of computing the mean (see p. 571), with zero at the assumed mean, those above running from + 1 up and those below from — 1 down. The method used in the example is recommended, however.

therefore Σx for this row is $5 \times 10 + 3 \times 11 + 1 \times 12 = 95$. In similar fashion Σx for each row is found. Each entry in that column is then multiplied by the corresponding d_y and the result placed under the heading Σxy. Thus $11 \times 11 = 121$, $10 \times 95 = 950$, and so on. In the table the same computations have been carried out in the other direction and the results entered in the Σy and Σxy rows at the bottom. These are not really necessary, however, but are useful inasmuch as they serve as a check upon the previous work, both yielding the same sum, in this case 4,316.

After obtaining the sums of the fd_x, fd_y, fd_x^2, fd_y^2, and Σxy series,[22] one may at once proceed to compute the coefficient of correlation by a formula that is really the same as the one given on page 582, but is commonly expressed in somewhat different symbols. It is

$$r = \frac{\Sigma xy - \frac{\Sigma fd_x \cdot \Sigma fd_y}{N}}{\sqrt{(\Sigma fd_x^2 - \frac{(\Sigma fd_x)^2}{N})\ (\Sigma fd_y^2 - \frac{(\Sigma fd_y)^2}{N})}}.$$

Substituting in this the sums of the proper columns,

$$r = \frac{4316 - \frac{594 \times 526}{80}}{\sqrt{(4822 - \frac{594^2}{80})\ (3924 - \frac{526^2}{80})}} = .94$$

for the two sets of eighty scores.

Rank correlation.—Although the product-moment method of correlation just described is generally considered the standard method, that of rank correlation is frequently employed when the number of cases is not over thirty or forty. It is based on the ranks of the scores or measures in the two series and does not take their exact sizes into account, therefore it is less exact than the former method. Since the product-moment correlation for a small number of cases is usually not very reliable, however, rank correlation may be considered as almost if not quite as accurate for small groups.

There are two common methods of computing rank correlation, both of which make use of ranks determined as shown on page 579.

[22] The sums of the Σx column and the Σy row need not be found, but may be used to check those of the fd_x row and the fd_y column, respectively.

TABLE XXVI. COMPUTATION OF RANK CORRELATION

X	Y	$Rank_x$	$Rank_y$		D	D^2
25	25	22	22		0	00
24	22	21	17	+	4	16.00
23	24	19.5	21	—	1.5	2.25
23	22	19.5	17	+	2.5	6.25
22	23	18	20	—	2.	4.00
21	22	17	17		0	.00
20	21	15	13.5	+	1.5	2.25
20	21	15	13.5	+	1.5	2.25
20	15	15	4	+	11.	121.00
19	22	12.5	17	—	4.5	20.25
19	20	12.5	12	+	.5	.25
18	17	11	7.5	+	3.5	12.25
17	22	9.5	17	—	7.5	56.25
17	19	9.5	10.5	—	1.	1.00
16	19	7	10.5	—	3.5	12.25
16	18	7	9	—	2.	4.00
16	13	7	2	+	5.	25.00
14	17	4	7.5	—	3.5	12.25
14	16	4	5.5	—	1.5	2.25
14	14	4	3	+	1.	1.00
13	16	2	5.5	—	3.5	12.25
9	12	1	1		0	.00
				$\Sigma g =$	30.5	313.00 $= \Sigma D^2$

$$R = 1 - \frac{6 \times 30.5}{22^3 - 1} = 1 - \frac{183}{483} = .62 -$$

$$\rho = 1 - \frac{6 \times 313}{22(22^2 - 1)} = 1 - \frac{1878}{10626} = .82 +$$

Probably the better of the two is that based upon the differences of the ranks and often called the "foot-rule formula." Its computation is somewhat simpler than that involved in the other, also on theoretical grounds it appears to have the advantage. To illustrate it the two series of twenty-two scores already employed for product-moment correlation will be used. As before, these are given in the first two, or X and Y, columns. It is convenient, but not at all necessary, to have the scores in one column in order. The next two columns contain the ranks of the X and Y scores, respectively. The next column, headed D, contains the differences between the ranks. These may be taken either by subtracting the X ranks from the Y ones, or vice versa. In the example the Y ranks have been subtracted from the X ones, as this is perhaps slightly less confusing than the reverse. Thus the first difference is 0, from 22 — 22; the second is + 4, from 21 — 17; the third is — 1.5, from 19.5 — 21; and so on. The sum of the positive differences or gains

is then determined, and may be checked by finding that of the negative differences, which is always the same. This sum, usually designated Σg (g for gain) is then substituted in the formula $R = 1 - \frac{6\Sigma g}{N^2 - 1}$, in which R is the symbol for rank correlation and N, of course, the number of cases. In the example given in the table, $\Sigma g = 25.5$, therefore $R = 1 - \frac{6 \times 30.5}{22^2 - 1} = 1 - \frac{183}{483} = .62 -$.

The other formula is based upon the squares of all the differences and is as follows, ρ (rho) being its symbol: $\rho = \frac{6\Sigma D^2}{N(N^2 - 1)}$ To illustrate its computation, the last or D^2 column of the table has been added. Its sum, 313, is substituted in the formula, along with $N = 22$, giving $\rho = 1 - \frac{6 \times 313}{22(22^2 - 1)} = 1 - \frac{1878}{10626} = .82 +$.

The reader is probably surprised to notice that the values of R and ρ, .62 and .82 respectively, differ so greatly. This is usual, since the basis of computation is different. Partly because of this difference, and partly because the product-moment coefficient is considered the standard, it is customary to change both R and ρ into approximately equivalent values of r, the product-moment coeffi-

TABLE XXVII. TABLE FOR CHANGING RANK CORRELATION R INTO PRODUCT-MOMENT CORRELATION r

R	r	R	r	R	r	R	r	R	r	R	r
.00	.000	.17	.291	.34	.541	.51	.742	.68	.889	.85	.975
.01	.018	.18	.307	.35	.554	.52	.753	.69	.896	.86	.979
.02	.036	.19	.323	.36	.567	.53	.763	.70	.902	.87	.981
.03	.054	.20	.338	.37	580	.54	.772	.71	.908	.88	.984
.04	.071	.21	.354	.38	.593	.55	.782	.72	.915	.89	.987
.05	.089	.22	.369	.39	.606	.56	.791	.73	.921	.90	.989
.06	.107	.23	.384	.40	.618	.57	.801	.74	.926	.91	.991
.07	.124	.24	.399	.41	.630	.58	.810	.75	.932	.92	.993
.08	.141	.25	.414	.42	.642	.59	.818	.76	.937	.93	.995
.09	.158	.26	.429	.43	.654	.60	.827	.77	.942	.94	.996
.10	.176	.27	.444	.44	.666	.61	.836	.78	.947	.95	.997
.11	.192	.28	.458	.45	.677	.62	.844	.79	.952	.96	.998
.12	.209	.29	.472	.46	.689	.63	.852	.80	.956	.97	.999
.13	.226	.30	.486	.47	.700	.64	.860	.81	.961	.98	.9996
.14	.242	.31	.500	.48	.711	.65	.867	.82	.965	.99	.9999
.15	.259	.32	.514	.49	.721	.66	.875	.83	.968	1.00	1.0000
.16	.275	.33	.528	.50	.732	.67	.882	.84	.972		

cient. These changes are made according to certain formulae, but in actual practice tables are employed, so that all one need do is to look up the obtained value of R or ρ and find the corresponding value of r. Doing so and looking up $R = .62$ in Table XXVII, one finds that the corresponding value of $r = .84$.

The other rank coefficient, ρ, differs much less from r than does R. Indeed, a table for it similar to that just given for R shows no difference quite as great as .02. Therefore no table for ρ and r is given here. Instead, the following rules may be followed: If ρ equals from .00 to .10, make no change; if from .11 to .36, add .01; from .37 to .76, add .02; from .77 to .94, add .01; from .95 to 1.00, do not change. Since $\rho = .82$ in the example, $r = .83$, just one point different from the value $r = .82$ secured through R. Both approximate quite closely the value of r, .83, already found by the product-moment method. This difference illustrates the statement made above that rank correlation methods yield only approximate values of r. In very few cases is the difference as great as .10 and in most it is not over .05.

Brown's formula. In connection with the use of the coefficient of correlation as a measure of reliability, in which connection it is commonly called the coefficient of reliability, there is frequently occasion to employ a formula commonly known as Brown's formula or as the Spearman-Brown formula. This is concerned with the relationship between the lengths of similar tests and their reliabilities. The formula is $r_n = \frac{Nr}{1 + (N-1)r}$, in which r_n is the coefficient of reliability of a similar test N times as long [24] as one for which the coefficient is r. For example, if a test has a reliability coefficient of .70, that of a similar test twice as long would be $\frac{2 \times .70}{1 + (2-1).70} = \frac{1.40}{1.70} = .82+$; that of one three times as long $\frac{3 \times .70}{1 + (3-1).70} = \frac{2.10}{2.40} = .88-$, and so forth. One of the most frequent applications of this formula is to determine the reliability of a whole test when scores on its odd items have been correlated with those on its even ones. In this case N, of course, equals two, since the whole test is twice as long as either half.

[24] Instead of a similiar test N times as long, one is occasionally concerned with the combined results from using the original test N times, which gives the same results in so far as reliability is concerned.

Sometimes it is convenient to reverse the process and determine how much longer a test must be to yield a certain coefficient of reliability when that secured from it is already known. The formula for this is $N = \frac{r_n(1-r)}{r(1-r_n)}$. For example, if a test has a coefficient of reliability of .75 and one of .90 is desired, $N = \frac{.90(1-.75)}{.75(1-.90)} = \frac{.225}{.075} = 3$, or, in other words, it must be three times as long.

The index of reliability.—The index of reliability is a rarely used measure, but one that has much to recommend it. Just as the coefficient of reliability is the coefficient of correlation between scores on two forms or applications of the same test, so the index of reliability is the coefficient between scores on one form or application and the theoretically true scores, which for practical purposes may be considered the averages of all available scores.[25] The method of computing the index of reliability is to find the coefficient of reliability and extract its square root. That is, index of reliability $= \sqrt{r}$. Thus if the coefficient of reliability is .64, for example, the index is .80; if the coefficient is .80, the index is .89 +; and so on.

The standard and probable errors of estimate and of measurement.—Just as reliability may be measured by the correlation between two sets of actual scores or by that between one such set and theoretically true scores, so also it may be measured in terms of the differences between such scores. The differences between actual scores resulting from different applications or forms are known as errors of estimate,[26] since they refer to errors present in estimating one series from the other. Those between actual and theoretically true scores are called errors of measurement, since they are the errors involved in using obtained scores as true measures.

The error of estimate or of measurement of any particular score

[25] A theoretically true score is defined as the average of an infinite number of actually obtained scores, corrected for practice effect and any other constant errors.

[26] Errors of estimate are not limited to differences between scores resulting from two applications of the same test or duplicate forms thereof, but may be found between any two correlated series of measures.

can rarely be determined, but the whole distribution of errors in any particular case can be described quantitatively. For this purpose the standard and probable errors are usually employed. The standard error is merely the standard deviation of the errors and the probable error the median deviation of the errors. These measures may, therefore, be interpreted as suggested on pages 578 and 579. About 68 per cent of the errors are smaller than the standard error, over 95 per cent smaller than twice the standard error, 99.7 per cent smaller than thrice it, and so on. Likewise 50 per cent are no larger than the probable error, over 82 per cent no larger than twice the probable error, almost 96 per cent no larger than three times it, and so on. Because of the fact that just 50 per cent of errors are smaller than the probable error and 50 per cent larger, or, in other words, that it is a "50–50" or even chance that any particular error is smaller or larger, whereas the standard error cannot be interpreted in terms so easy to remember, the probable error has come into much more common use than the other.

If the standard deviations and coefficient of reliability of two sets of scores are known, the computation of the standard and probable errors of estimate and of measurement is easy. The formulae are as follows:[37]

$$\sigma_{est.} = \sigma\sqrt{1 - r^2} \qquad \text{P.E.}_{est.} = .6745\sigma\sqrt{1 - r^2}$$

$$\sigma_{meas.} = \frac{\sigma_x + \sigma_y}{2}\sqrt{1 - r} \qquad \text{P.E.}_{meas.} = .6745\frac{\sigma_x + \sigma_y}{2}\sqrt{1 - r}.$$

Thus to secure the standard error of measurement, all that need be done is to multiply the standard deviation by the square root of the radical one minus the square of the coefficient of reliability. The standard deviation used is that of the series of scores being estimated or dealt with in terms of the other. For example, if algebra and geometry scores have been correlated and the latter are being estimated in terms of those in algebra, their standard deviation is used. In the formula for the standard error of measurement, the average of the standard deviations of the two series

[37] There are several different formulae for the standard and probable errors of measurement, yielding results which vary somewhat in their significance. The ones given above are, however, far more often used than any of the others.

is employed and the quantity under the radical is $1 - r$ instead of $1 - r^2$. In both cases the probable errors are found by multiplying the standard errors by .6745, just as the median deviation is obtained from the standard deviation.

In connection with errors of estimate the coefficient of alienation should be mentioned. This expression, abbreviated k, equals $\sqrt{1 - r^2}$, and thus varies from 0 to 1.00 inversely with r. It is sometimes used alone as a measure of the departure from perfect correlation or reliability, or of the size of errors involved in estimating one series of scores from another. Probably the best interpretation of the coefficient of alienation is that it measures the size or proportion of the guessing element present in estimating one series of measures from another with which it has the given correlation. Thus if $r = 1.00$, $k = .00$, so there is no guessing at all but absolute certainty of prediction. If $r = .95$, $k = .31 +$ so the element of guess is almost one-third, or, in other words, the errors in the estimated scores are almost one-third as large as they would be in pure guesses. If $r = .90$, $k = .44 -$, or errors more than three-sevenths as large as those in pure guesses are present. Similarly, if $r = .87 -$, the guessing element is one-half; if $r = .80$ it is three-fifths; if $r = .60$ it is four-fifths and so on down until if $r = .00$ it is one, or an entire guess. Since $\sqrt{1 - r^2}$ appears in the formula for the standard error of estimate this formula may also be written $\sigma_{\text{est.}} = \sigma k$, but this form is rarely encountered.

Sometimes one further step is taken and a measure of the element of certainty of prediction secured by subtracting the coefficient of alienation from one, the result being called the efficiency of prediction. In formula form $E = 1 - k$. Thus if $r = .95$ and k correspondingly $= .31$, $E = .69$, and so on for other values.

As was stated on page 63, the direct comparison of errors of measurement does not possess great significance. The reason is that their size depends on other factors than reliability, especially the size and spread of the scores themselves. Therefore it is better to employ the ratio of the probable error to some quantity representative of these latter factors. Two, the mean and the standard deviation, have received frequent use. Each has certain advantages for this purpose, so that it is better to use both than either

one alone. Therefore the recommended measures of reliability from this standpoint are $\frac{\text{P.E.}_{\text{meas.}}}{M}$ and $\frac{\text{P.E.}_{\text{meas.}}}{\sigma}$.[28]

Regression.—Several references have been made to estimating scores in one series from those in another correlated with them and to the errors involved in such estimates, but nothing has been said as to just how the estimates are made. Before giving the method, it seems well to explain briefly the circumstances under which estimates are made. For series of scores actually correlated they are not ordinarily made, since the scores in both series must be already known in order to compute the correlation and so there is usually no point to making estimates. In situations similar to one in which the correlation has been found it is frequently the case that one series is known and there is a reason for predicting the other. It is then assumed that the same relationship holds as was found for the two similar series actually correlated and on this assumption the unknown measures are estimated from the known ones. Thus, for example, if the freshman algebra marks and the sophomore geometry marks of the same pupils have been correlated, there is no occasion for estimating one from the other for the pupils concerned. For another similar group of pupils whose algebra marks were known and who had not yet begun geometry it might be desirable to estimate or predict their geometry marks for purposes of guidance, grouping, or something else. It would then be assumed that the relationship between algebra and geometry marks was the same in the second group as in the first and, therefore, the predictions made according to the data obtained from its marks.

For the purpose of estimating or predicting measures in one series from those in another, one must know the means and standard deviations of both series and the coefficient of correlation between the two series. The regression equations, which are those used in making estimates, are as follows, using X to represent one variable and Y to represent the other: $X = r\frac{\sigma_x}{\sigma_y}Y + M_x -$

[28] Since $\text{P.E.}_{\text{meas.}} = .6745\frac{\sigma_x + \sigma_y}{2}\sqrt{1 - r^2}$ and σ_x and σ_y may be considered equal in the case of duplicate forms, $\frac{\text{P.E.}_{\text{meas.}}}{\sigma} = .6745\sqrt{1 - r^2}$ or $.6745\,k$.

$r\frac{\sigma_x}{\sigma_y}M_y$ and $Y = r\frac{\sigma_y}{\sigma_x}X + M_y - r\frac{\sigma_y}{\sigma_x}M_x$.[29] The first or X equation is employed to estimate X scores when Y scores are known and the second or Y equation to estimate Y scores when X scores are known. As will be seen in the next paragraph, these equations reduce to a very simple form when the numerical values of r, the σ's and M's are substituted in them.

The use of regression equations may be illustrated in connection with the two series of eighty scores each employed for the computation of the coefficient of correlation on page 587. Using X for the first scores and Y for the second ones, $M_x = 74.63$, $M_y = 75.38$, $\sigma_x = 11.34$, $\sigma_y = 12.06$ and $r = .938$.[30] Substituting these values in the two equations gives $X = .938\frac{11.34}{12.06}Y + 74.63 - .938\frac{11.34}{12.06}75.38$ and $Y = .938\frac{12.06}{11.34}X + 75.38 - .938\frac{12.06}{11.34}74.63$. These reduce to $X = .882Y + 74.63 - .882 \times 75.38 = .882Y + 8.14$ and $Y = .998X + 75.38 - .998 \times 74.63 = .998X + .90$. Thus all that need be done to use the equations is to multiply the given value of Y by .882 and add 8.14 to estimate X and to multiply that of X by .998 and add .90 to get Y. Thus if a pupil has a score of 80 on the first or X series, his most likely Y score is $.998 \times 80 + .90 = 80.74$; if another has an X score of 55 his Y score is most likely $.998 \times 55 + .90 = 55.79$, and so on. Similarly if a pupil has a Y score of 90, for example, the best estimate of his X score is $.882 \times .90 + 8.14 = 87.52$. The differences between these estimated scores and those actually made are errors of estimate.

The unreliability of sampling.—Just as test scores are not perfectly reliable because they are not complete and accurate measures of the ability or trait in question, so all measures are to a certain extent unreliable when considered as representative of all cases of the kind upon which they are based. For example,

[29] The expressions "$r\frac{\sigma_x}{\sigma_y}$" and "$r\frac{\sigma_y}{\sigma_x}$" which occur in these equations are known as the regression coefficients and are abbreviated b_x and b_y, respectively. Using them, the equations may be written: $X = b_xY + M_x - b_xM_y$ and $Y = b_yX + M_y - b_yM_x$.

[30] Usually r should be carried to three, or perhaps four, decimal places when it is to be used in a regression equation.

the mean of the twenty-two vocabulary test scores is perfectly reliable only when considered as the mean of those particular scores, not as the mean of another similar set of twenty-two scores made by the same pupils or as that of all similar pupils. It is, however, very commonly convenient or even necessary to take an average, a measure of variability, a coefficient of correlation or some other measure as representative of a much larger group of pupils than that used in computing it. When this is done it is usually desirable to accompany the measure with some information as to its reliability. Ordinarily this is done by stating either its standard or its probable error, the latter being much more common. The meaning of these is the same as of the standard and median deviations in general. That is, there are 68 + chances out of one hundred that the obtained measure differs from the true one—that for the total larger group—by not over one standard error, 95 + chances out of one hundred that it differs by not over twice the standard error, and so on. Similarly there are 50 chances out of one hundred that it does not differ by more than one probable error, 82 + out of one hundred that it does not differ by more than twice the probable error, and so on. It is customary to connect a probable error with the measure to which it applies by a plus-or-minus sign ($\pm$). Thus, for example, $M = 82 \pm 2.5$ is equivalent to stating that the mean is 82 with an even chance that it does not differ over 2.5 from the mean of the whole group of which only a part was used in the computation; $r = .72 \pm .06$ means that the coefficient of correlation actually found is .72 and the chances even that it differs no more than .06 from that of all similar cases.

As would naturally be expected, the larger the sample selected from the whole group or population, as it is technically called, the greater the reliability of measures computed from the sample when considered as representative of the whole population. The increase in reliability is in direct ratio to the square root of the increase in the size of the sample. Thus measures for a sample four times as large as another will be twice as reliable, those for a sample nine times as large, thrice as reliable, and so on. Since the reliability increases with the number of cases, the standard and probable errors naturally decrease.

In most cases the formulae for determining errors are simple

and easy to employ. Those for both the standard and the probable error of a number of the most commonly used measures are given in Table XXVIII.

TABLE XXVIII. FORMULAE FOR THE STANDARD AND PROBABLE ERRORS OF CERTAIN COMMONLY USED MEASURES *

Measure	*Standard error*	*Probable error* †
Mean	$\frac{\sigma}{\sqrt{N}}$	$.67\frac{\sigma}{\sqrt{N}}$ or $\frac{\text{Md. D.}}{\sqrt{N}}$
Median	$1.25\frac{\sigma}{\sqrt{N}}$	$.85\frac{\sigma}{\sqrt{N}}$ or $1.25\frac{\text{Md. D.}}{\sqrt{N}}$
First or third quartile	$1.36\frac{\sigma}{\sqrt{N}}$	$.92\frac{\sigma}{\sqrt{N}}$ or $1.36\frac{\text{Md. D.}}{\sqrt{N}}$
Quartile deviation	$.79\frac{\sigma}{\sqrt{N}}$	$.53\frac{\sigma}{\sqrt{N}}$ or $.79\frac{\text{Md. D.}}{\sqrt{N}}$
Standard deviation	$.71\frac{\sigma}{\sqrt{N}}$	$.48\frac{\sigma}{\sqrt{N}}$ or $.71\frac{\text{Md. D.}}{\sqrt{N}}$
Median deviation	$.48\frac{\sigma}{\sqrt{N}}$	$.32\frac{\sigma}{\sqrt{N}}$ or $.48\frac{\text{Md. D.}}{\sqrt{N}}$
Coefficient of correlation	$\frac{1-r^2}{\sqrt{N}}$	$.67\frac{1-r^2}{\sqrt{N}}$

The decimals given are only approximate, being carried to just two places, but are exact enough for ordinary use. It will be noted that the expression $\sqrt{N}$ always appears in the denominator, thus expressing mathematically the fact that errors of sampling decrease in inverse ratio to the square root of the number of cases. Also it is apparent that the only quantities needed in order to determine the standard or probable errors of the measures listed are, except in the last case, the standard deviation and the number of cases.

To illustrate the use of these formulae and indicate something of the size of errors, reference will be made to the series of eighty scores first given on page 556. Their standard deviation is 11.34, hence the standard error of the mean, commonly abbreviated σ_m, is $\frac{11.34}{\sqrt{80}} = 1.27$ and its probable error .85. In other words the

* These formulae are merely approximate, the decimals being given to only two places.

† Two formulae for most of the probable errors are given, since sometimes it is more convenient to employ one and sometimes the other.

chances are even that the mean, 74.625, is not more than .85 different from the mean of all similar pupils on the same test. For the median, which is 75.42, $\sigma_{Md.} = 1.25\frac{11.34}{\sqrt{N}} = 1.59$ and P.E.$_{Md.} = 1.06$. For the coefficient of correlation between the two sets of eighty scores found on page 586, P.E.$_r = .67\frac{1 - .94^2}{\sqrt{80}} =$ .01. Therefore the coefficient of correlation for these two series may be written as $.94 \pm .01$.

Summary.—In this chapter a number of the statistical methods most often employed in handling test scores have been treated, the necessary formulae given, the computation of the measures illustrated and some brief interpretations given. The topics dealt with include tabulation and classification, the mean, median, and mid-score, the first and third quartiles, percentiles, the quartile, standard and median deviations, correlation, both product-moment and rank, Brown's formula, the index of reliability, errors of estimate and of measurement, regression and errors of sampling.

BIBLIOGRAPHY[81]

Barlow's *Tables of Squares, Cubes, Square Roots, Cube Roots, Reciprocals, of All Integer Numbers up to 10,000*, Stereotype Edition. New York: Spon and Chamberlain, 1914. 200 p.

Garrett, H. E. *Statistics in Psychology and Education*. New York: Longmans, Green and Company, 1926. 317 p.

Greene, H. A. *Work-Book in Educational Measurements*. New York: Longmans, Green and Company, 1928. 156 p.

Gregory, C. A. and Renfrow, O. W. *Statistical Method in Education and Psychology*. Cincinnati, Ohio: C. A. Gregory Company, 1929. 228 p.

Holzinger, K. J. *Statistical Methods for Students in Education*. Boston: Ginn and Company, 1928. 372 p.

———. *Statistical Tables for Students in Education and Psychology*. Chicago: University of Chicago Press, 1925. 74 p.

Holzinger, K. J. and Mitchell, B. C. *Exercise Manual in Statistics*. Boston: Ginn and Company, 1929. 160 p.

Kelley, T. L. *Statistical Method*. New York: The Macmillan Company, 1923. 390 p.

[81] This list of references includes not only texts and sets of exercises, but also a few tables that are useful in computation.

Lindquist, E. F. and Stoddard, G. D. *Study Manual in Elementary Statistics.* New York: Longmans, Green and Company, 1929. 109 p.

Morton, R. L. *Laboratory Exercises in Educational Statistics, With Tables.* New York: Silver, Burdett and Company, 1928. 197 p.

Odell, C. W. *Educational Statistics.* New York: The Century Co., 1925. 334 p.

Otis, A. S. *Statistical Method in Educational Measurement.* Yonkers, New York: World Book Company, 1925. 337 p.

Rugg, Harold. *A Primer of Graphics and Statistics for Teachers.* Boston: Houghton Mifflin Company, 1925. 142 p.

Symonds, P. M. "How to Handle the Results of Testing," *Measurement in Secondary Education.* New York: The Macmillan Company, 1927, Chapter XII.

Thurstone, L. L. *The Fundamentals of Statistics.* New York: The Macmillan Company, 1925. 237 p.

Trabue, M. R. *Measuring Results in Education.* New York: American Book Company, 1924, Chapters V, IX, X, XI, XVII.

Whitney, F. L. *Statistics for Beginners in Education.* New York: D. Appleton and Company, 1929. 123 p.

Williams, J. H. *Elementary Statistics.* New York: D. C. Heath and Company, 1929. 220 p.

CHAPTER XXV

GRAPHS

The normal frequency curve.—Of the various types of curves and graphs employed and referred to in connection with test scores and other measures, the normal frequency curve [1] is by far the most commonly mentioned and employed. It is a symmetrical bell-shaped curve, highest at its center, descending rather sharply close to its peak and then less and less so as the distance from the center increases. Theoretically it never touches the base line, but it comes so close to it that the intervening space cannot be shown at a distance greater than three or four standard deviations from the center except on very large scale drawings. Its center, or highest point, always coincides with the mean and median of the normal distribution it represents and the point at which the slope of the curve becomes more nearly horizonal than vertical [2] is 1σ from the center. Figure 2 contains representations of two normal curves, drawn on the same horizontal scale, but with different vertical scales, that of the upper being twice that of the lower. Both are cut off at 3σ from the mean, or center, in each direction. In drawing the normal curve, or any other, the horizontal and vertical scales used are matters of preference and their exact relationship to each other is usually not highly important. For most purposes, however, the writer suggests that the scales used be such that the greatest height is from half to all the width, probably nearer the first. The upper one of the two curves is, therefore, better proportioned than the lower, which is too low for its width. In the normal curve as in most others the vertical or y-distance represents the number of cases and the horizontal or x-distance the size of the scores or measures.

Although the normal curve has a definite mathematical basis,

[1] This is also known as the normal probability curve, the curve of error and by other names.

[2] This is called the point of inflection. At it the curve makes an angle of 45° with the base, and of course the same with a vertical line through the center.

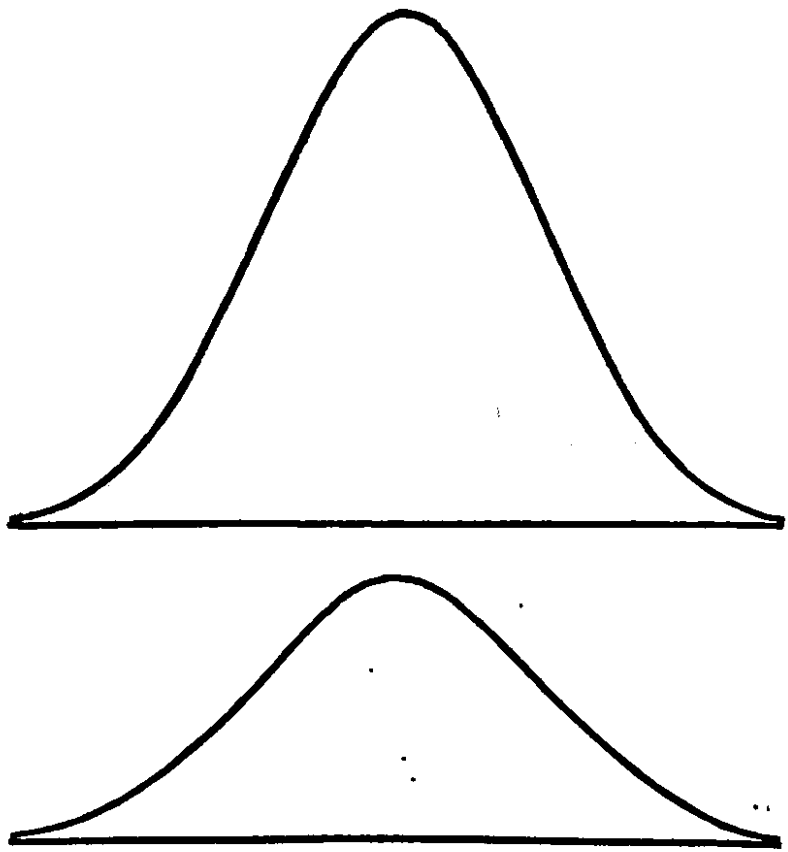


FIGURE 2
NORMAL CURVES

being the graphic representation of the formula,[3] also the curve

$$y = y_0 e^{\frac{-x^2}{2\sigma^2}}$$

[3] In this formula x and y are used, as commonly in mathematics, to represent horizontal and vertical distances, respectively. In more technical language, x is an abscissa and y an ordinate. y_0 is the height of the curve at its center or highest point, that is, the maximum ordinate, and may be found in any particular case by the formula $y_0 = \frac{N}{\sigma\sqrt{2\pi}}$; e is the base used in calculating logarithms, 2.7183 —.

obtained from the binomial expansion when the two terms are equal, its chief educational importance is not based directly thereon. It is rather due to the fact that if most biological and many other traits or characteristics are measured for total populations or fairly large random samples thereof and the results graphed, they approximate such a curve very closely. The lengths of leaves of a plant, the weights of children of the same age, the test scores made by pupils in a single grade, may be cited as examples. So nearly universal is the normal curve that it is a fairly safe assumption that the results from the measurement of any mental or physical ability or trait of a homogeneous group will not vary far from it. Moreover, it has been found that chance or variable errors, such as those in shooting at a target, in making physical or mental measurements, and so forth, likewise approximate the normal curve. A third though less important fact is that operations in which there are equal chances, such as tossing coins to determine heads and tails, drawing balls from a container holding an equal number of each of two kinds, and so forth, yield similar results. The underlying cause of all these facts is that the normal curve or distribution results from the operation of a large number of minute causes and that this situation holds in all the cases just mentioned.

There are two chief uses of the normal curve in the field of educational measurements. One of these is that many statistical measures, such as the probable error, the standard deviation or error, the coefficient of correlation and many others, are commonly employed and interpreted on the assumption that all distributions in connection with which they are used are normal. Unless distributions differ considerably from normal, however, the errors involved in the assumption of normality are usually not very great. It has, therefore, become customary to assume it in handling test scores and other more or less similar data.

The other use referred to is that actually obtained distributions of test scores or other measures are compared with the normal to ascertain if the sample dealt with is random, the measuring instrument satisfactory as to difficulty and scaling, and so forth. The complete procedure in making such comparisons involves considerable calculation and is rarely worth the while of the regular teacher, so will not be presented here. Instead of employ-

ing the method referred to teachers and others are advised merely to graph the distribution in question and, on the same figure, the best fitting normal curve, and compare the two. If no large differences exist, one may conclude that the distribution is, for all practical purposes, normal.

Skew curves.—A skew curve may be thought of as a normal curve that has been pushed or pulled in one direction. If it has been pushed in from the left, or pulled out to the right, the curve is said to be positively skewed or to have plus skewness. The two curves at the right in Figure 3 illustrate this. If the curve has

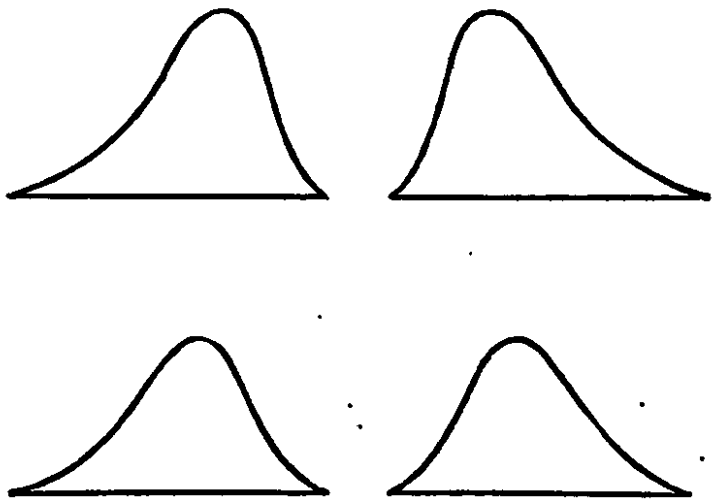


FIGURE 3
SKEW CURVES

been pushed in from the right or pulled out to the left, the skewness is negative or minus. The two curves at the left in Figure 3 are negatively skewed.

The situations that most commonly yield results forming a skew curve are those in which some biased or non-random selection has taken place. For example, if instead of testing a random group of eight-year-old children one tests only those in the fourth grade, who are usually rather highly selected as to ability, the resulting distribution of scores will probably be skewed upward

or positively and somewhat resemble one of the curves at the right in Figure 3. If, on the other hand, only the eight-year-olds in Grades I and II are measured, the curve representing their scores will probably be negatively skewed and resemble one of those at the left in Figure 3.

The histogram or column diagram.—Three varieties of the ordinary frequency curve are more or less commonly employed to represent distributions of scores. These are the histogram or column diagram, the frequency polygon, and the smooth frequency curve. Of these the first is recommended as generally to be preferred, chiefly because it is a more exact portrayal of the particular scores dealt with and, therefore, more easily understood and interpreted. The histogram is composed of a series of rectangles, each of which has as its base one class interval and as its height the number of cases in the particular interval it represents. Such a curve is shown in Figure 4. The distribution

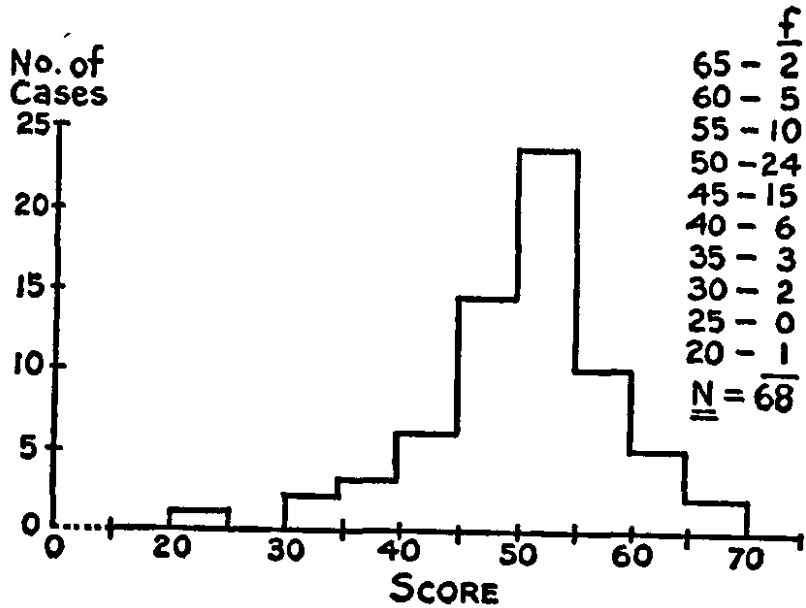


FIGURE 4

HISTOGRAM OR COLUMN DIAGRAM REPRESENTING DISTRIBUTION OF SCORES ON AN INTELLIGENCE TEST

graphed, given at the right of the figure, is of the scores made by sixty-eight pupils on an intelligence test. Beginning at the left, the first rectangle, based on the interval from 20 up to 25, has a height of one, the number of cases therein. There is none for the next interval, there the frequency is zero. The rectangle above the interval from 30 to 35 is two units high, that above the next three units, and so on, as indicated by the frequencies. Broken vertical lines have been employed between adjacent rectangles. Sometimes solid ones are used, sometimes none at all, according to whether it is desired to emphasize the separate classes or the distribution as a whole.

It will be noted that at the left end of the base line there is a short broken line from 0 to 20. It is broken rather than solid to indicate that the distance between those points is not so long as it should be according to the scale used. Such a broken line is regularly employed when the distance from the lower limit of the first class to the zero point is greater than one or two intervals.

The histogram or column diagram lends itself very readily to the representation of the individual scores of a class or other group of pupils, so that each can see graphically how he compares with the others. This is accomplished by dividing each rectangle into sections, usually squares, one for each pupil, and labelling each with an identification mark.[4] Figure 5 presents such a graph

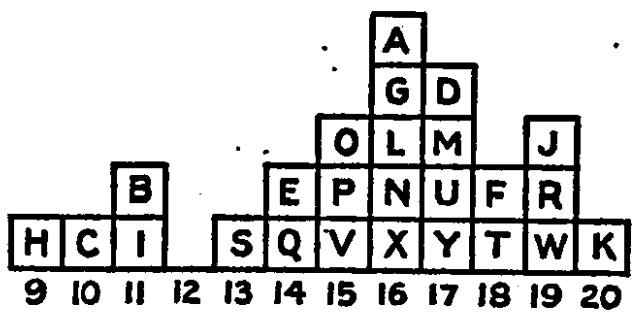


FIGURE 5

HISTOGRAM OR COLUMN DIAGRAM REPRESENTING INDIVIDUAL SCORES ON AN ALGEBRA TEST

[4] It is probably best to use identification marks known only to the pupils to whom they refer and to the teacher, so that the score of each pupil is not known by all the others.

for the results of a twenty-problem test given an algebra section of twenty-five pupils. The letters *A* to *Y*, inclusive, are used to designate the various pupils. Thus it shows pupil *H* that he solved only nine problems correctly and ranked at the very bottom, pupil *C* that he solved ten and ranked next to lowest, pupils *A*, *G*, *L*, *N*, and *X* that they solved sixteen and were at the middle of the class, pupil *K* that he alone solved all twenty correctly, and so on for the others. Such a graph is quickly made after one has acquired a little practice and makes very clear the comparative as well as the absolute standing of each pupil.

The frequency polygon.—A frequency polygon representing the same data as the histogram in Figure 4 is shown in Figure 6.

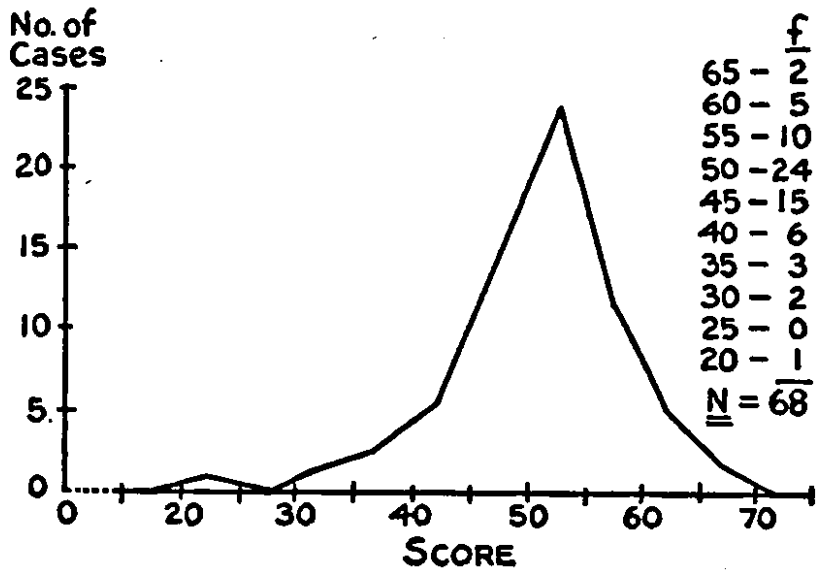


FIGURE 6

FREQUENCY POLYGON REPRESENTING THE SAME DISTRIBUTION OF SCORES AS THE HISTOGRAM IN FIGURE 4

From this it can be seen that the polygon is constructed by joining, with straight lines, points above the mid-points of the intervals located at heights equal to the numbers of cases in the various intervals. Thus, beginning at the left, the first of the points re-

ferred to is at a height of one above 22.5, the mid-point of the 20– class; the next is at a height of zero, that is, on the base line, at 27.5, the mid-point of the 25– class; the next is at a height of two above 32.5, and so on. At the ends the two extreme points, those above 22.5 and 67.5, are connected with the base line at the mid-points of the intervals just above and below them, respectively, these being 17.5 and 72.5. This practice is always to be followed, as the ends of the frequency polygon must not be left "up in the air."

Although the writer recommends the histogram in preference to the polygon as being more easily understood, the latter has some advantages. If the distribution is composed of classes that include more than one score each, as is usually the case, it probably approximates the actual distribution more nearly than does the histogram. The same is also true if it is considered to represent the probable distribution of a larger group of which only a sample has been taken rather than only the cases actually measured.

The smooth frequency curve.—This curve is more like the frequency polygon than the column diagram, being determined by points located in the same manner as for it. In other words, for the same data the same points are used for the two curves. Through these points a smooth line is drawn, usually merely by skill of hand and eye although so-called French curves"[5] may be used to assist in the process. A smooth curve for the same data represented by the histogram and polygon may be found in Figure 7.

The smooth frequency curve possesses in a still greater degree the two advantages which the frequency polygon has over the histogram. It is not here recommended for frequent use by the teacher, however, for the same reason that the polygon was not, and also because it is somewhat more difficult to draw well.

The cumulative frequency curve.—This curve differs from the three types just described in that its height at any given point indicates the total number of cases below or up to that point, if it rises toward the right, as is generally the case, or at or above that point, if it rises toward the left. It is usually drawn as a smooth

[5] These are curved figures of various shapes, frequently made of some transparent or semi-transparent material, which may be employed to assist one in drawing a smooth curve through two or more points.

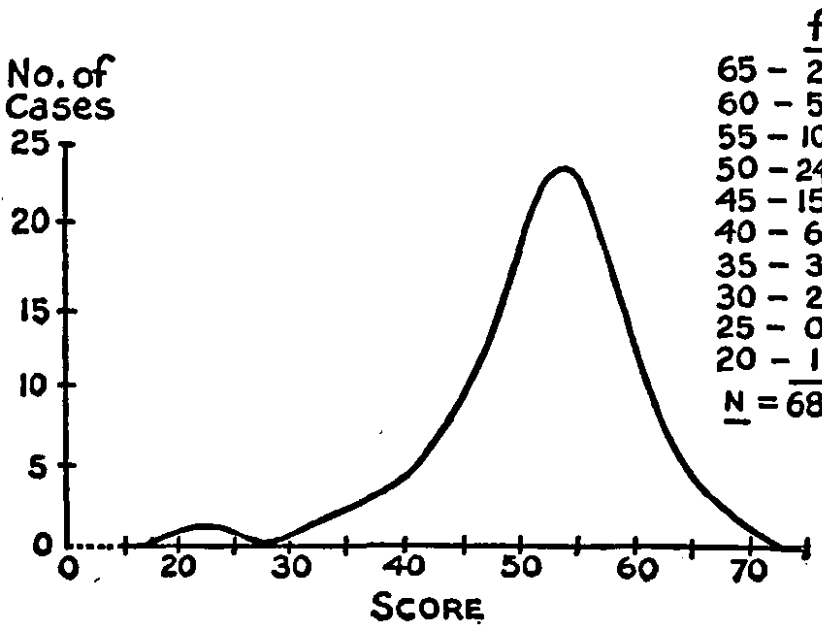


FIGURE 7

SMOOTH CURVE REPRESENTING THE SAME SCORES AS THE HISTOGRAM IN FIGURE 4 AND THE POLYGON IN FIGURE 6.

curve, although occasionally it is similar to a histogram and may be called a "stair-step" curve. Figure 8 illustrates it by presenting a smooth cumulative frequency curve for the same sixty-eight measures represented by the other curves. It starts from the base line at 20, since that is the lower limit of the whole distribution, or point below which there are no cases at all. At 25 it has a height of one, as there is a single case, that in the 20– class, below that point. At 30 its height is still one, as there are no additional cases below that point. At 35 it is three, there being also the two cases in the 30– class below 35; at 40 it is six, and so on. It ends at 70, the upper limit of the distribution, at a height of sixty-eight, the total number of cases.

The other type of cumulative frequency curve, that shows the number of cases at or above each point, for the same data is shown in Figure 9. It rises from the base line at 70, since no cases

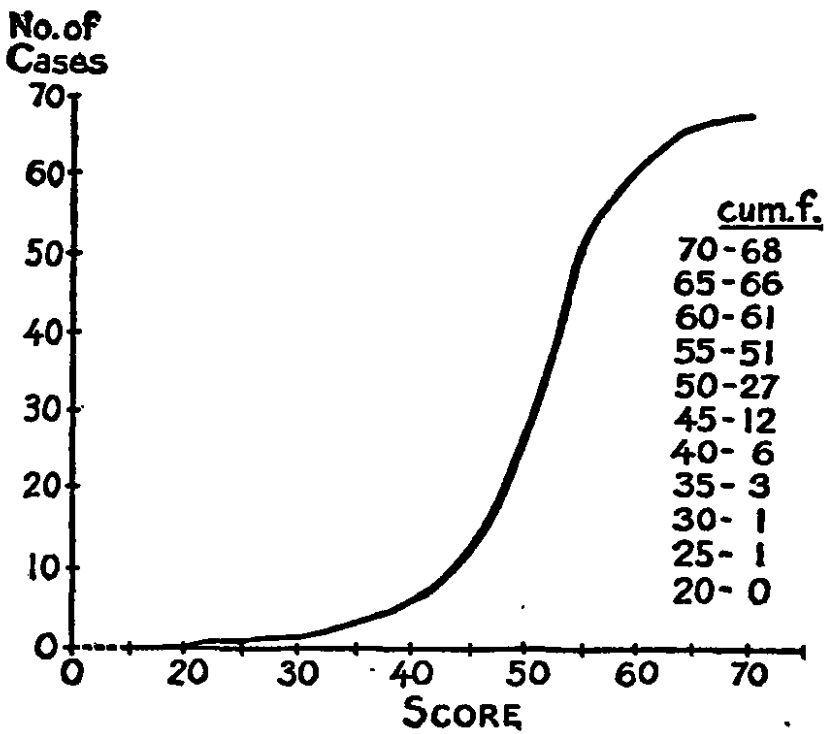


FIGURE 8

CUMULATIVE FREQUENCY CURVE REPRESENTING THE SAME SCORES AS THE CURVES IN FIGURES 4, 6, AND 7

are there or above, to a height of two at 65, of seven at 60, and so on to one of sixty-eight at 20, the point at or above which all the cases lie.

The ogive or percentile curve.—Although these terms are sometimes used interchangeably with cumulative frequency curve, it is best to restrict them to a somewhat similar but yet different variety of curve. The one point of difference is that the curve is drawn with scales interchanged on the axes. In other words, the scale of measurement is laid off on the vertical or Y-axis and the number of cases on the horizontal or X-axis. This type of curve, beginning at the lower left and rising to the right,

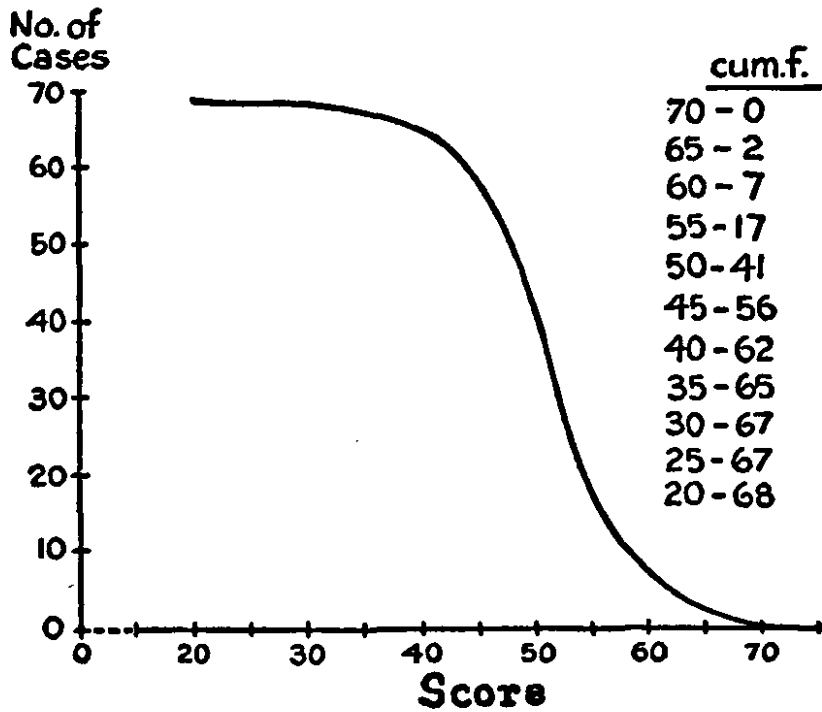


FIGURE 9

CUMULATIVE FREQUENCY CURVE REPRESENTING THE SAME SCORES AS THE CURVES IN FIGURES 4, 6, AND 7

has come to be used frequently in connection with test norms. Such a curve for the same data as the others is given in Figure 10. The reader will readily see that it may be interpreted in the same way as that in Figure 8, that there is one case below 25, also just one below 30, three below 35, and so on.

The reader may wonder why such a curve as that in Figure 10 is called a percentile curve, since no percentiles appear on that figure. The reason is that it is very common to draw vertical lines representing a number of the percentile points. The height or score value of the points at which these lines intersect the curve are the percentile points for the distribution represented by the curve. The next figure, Number 11, is the same as the last except

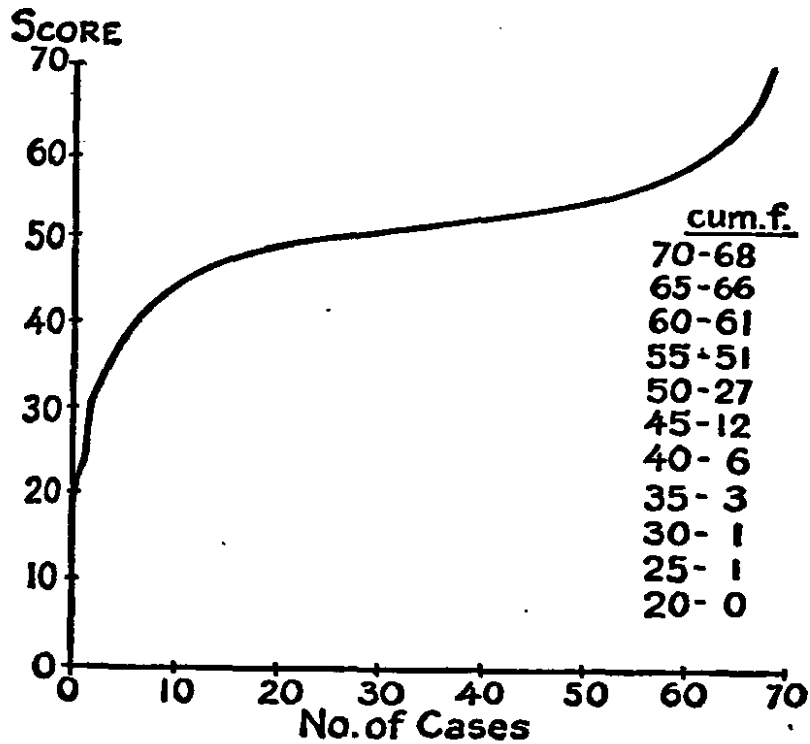


FIGURE 10

OGIVE OF PERCENTILE CURVE REPRESENTING THE SAME SCORES AS THE CURVES IN THE LAST TWO FIGURES

that vertical lines have been drawn at the fifth, tenth, twenty-fifth, fiftieth, seventy-fifth, ninetieth and ninety-fifth percentiles. Also light horizontal lines have been drawn, merely to aid in determining the height of the curve, or score, at the percentile points. From the figure one can see that the fifth percentile is about 36, the tenth about 41, the twenty-fifth or first quartile about 47, the fiftieth or median about 51, and so on.

The reader will note that there is a double scale at the bottom of the figure, showing both the actual frequencies and the per cents. In percentile graphs accompanying standardized tests the actual frequencies are rarely given, but only the percentile points.

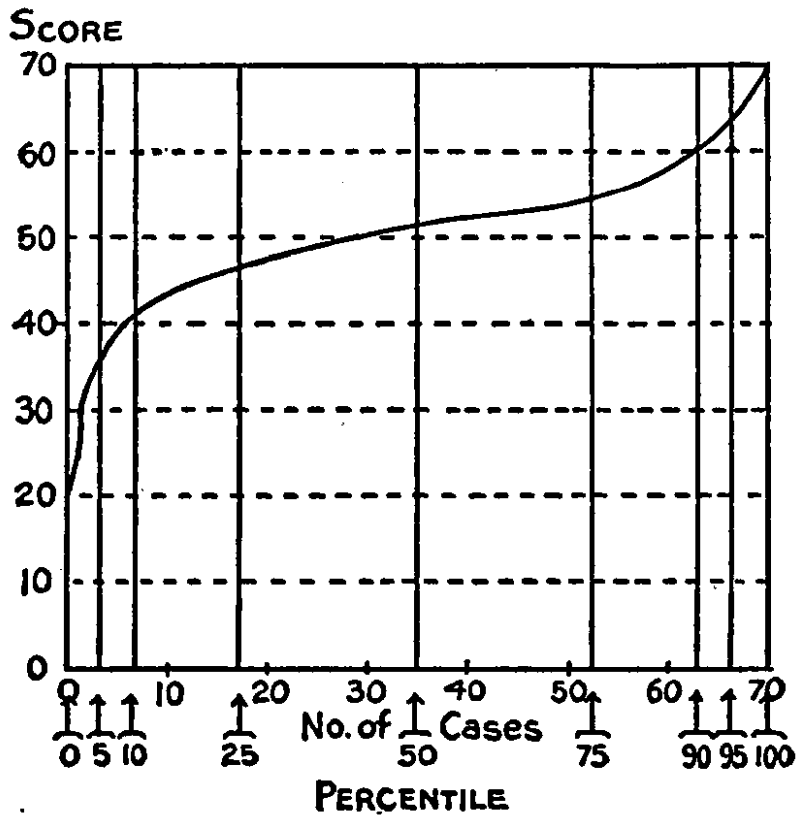


FIGURE 11

SAME OGIVE OR PERCENTILE CURVE AS IN FIGURE 10 WITH CERTAIN PERCENTILE LINES ADDED

The teacher or other user of standardized tests can, of course, compare the results from the group of pupils tested with those for pupils in general as shown on a published percentile curve by merely comparing a few percentile points. A more complete picture of the situation can be obtained by actually drawing in the percentile curve of the scores made for comparison with the given curve. Figure 12 illustrates this procedure. The heavy line represents the distribution of scores of about twelve thousand high

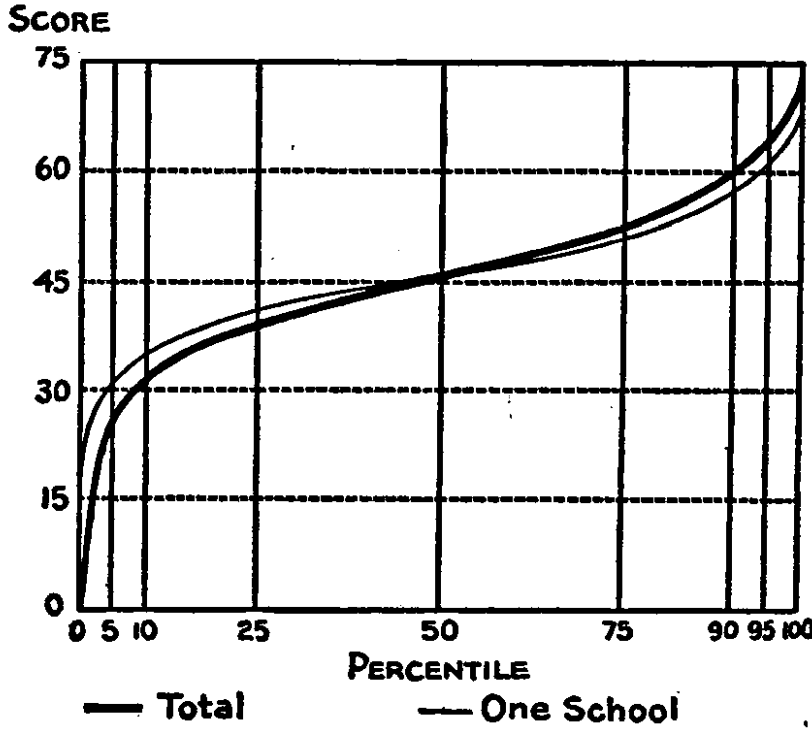


FIGURE 12

PERCENTILE GRAPH SHOWING TOTAL DISTRIBUTION OF SCORES AND DISTRIBUTION OF THOSE FROM ONE HIGH SCHOOL

school seniors on an intelligence test[6] and the light line that of the seniors in one high school.

It will be seen that the curve for the single school begins at a higher point on the vertical scale than does the heavy curve and ends at a lower one, but in its middle portion differs from it by very little. This shows that the scores made by seniors in the school in question averaged very nearly the same as those for the whole group from all schools, but that none of them were as low as the lowest or as high as the highest. In other words, the

[6] *See* Odell, C. W. "Conservation of Intelligence in Illinois High Schools," *University of Illinois Bulletin*, Vol. 22, No. 25, Bureau of Educational Research Bulletin No. 22. Urbana: University of Illinois, 1925. 55 p.

least intelligent seniors in the one school were more intelligent, and the most intelligent less so, than was true of schools in general, whereas those not near the extremes showed very little difference therefrom.

Summary.—In this chapter the writer has attempted to describe and illustrate briefly the types of curves most commonly employed and encountered in connection with educational measurements. These include the normal frequency curve, skew curves, the histogram or column diagram, the frequency polygon, the smooth frequency curve, cumulative frequency curves, and the ogive or percentile curve.

BIBLIOGRAPHY

Alexander, Carter. *School Statistics and Publicity.* New York: Silver, Burdett and Company, 1919, Chapters IV and XI.

Brinton, W. C. *Graphic Methods for Presenting Facts.* New York: *The Engineering Magazine*, 1914. 371 p.

Garrett, H. E. "Graphic Methods and the Normal Curve," *Statistics in Psychology and Education.* New York: Longmans, Green and Company, 1926, Chapter II.

Gregory, C. A. and Renfrow, O. W. "Graphical Presentation of Statistical Data," *Statistical Method in Education and Psychology.* Cincinnati: C. A. Gregory Company, 1929, Chapter IV.

Holzinger, K. J. *Statistical Methods for Students in Education.* Boston: Ginn and Company, 1928, p. 31–46, 131–36, 204–30.

Holzinger, K. J. and Mitchell, B. C. *Exercise Manual in Statistics.* Boston: Ginn and Company, 1929, p. 32–40, 70–87.

Kelley, T. L. "Graphic Methods," *Statistical Method.* New York: The Macmillan Company, 1923, Chapter II.

Lindquist, E. F. and Stoddard, G. D. *Study Manual in Elementary Statistics.* New York: Longmans, Green and Company, 1929, p. 30–41.

Odell, C. W. "Graphs," *Educational Statistics.* New York: The Century Co., 1925, Chapter II.

Otis, A. S. *Statistical Method in Educational Measurement.* Yonkers, New York: World Book Company, 1925, p. 30–35, 43–48, 53–84.

Rugg, Harold. *A Primer of Graphics and Statistics for Teachers.* Boston: Houghton Mifflin Company, 1925. 142 p.

Thurstone, L. L. *The Fundamentals of Statistics.* New York: The Macmillan Company, 1925, p. 9–17, 47–50, 143–54.

Williams, J. H. "Graphic Methods," *Elementary Statistics.* New York: D. C. Heath and Company, 1929, Part V.

Williams, J. H. *Graphic Methods in Education.* Boston: Houghton Mifflin Company, 1924. 319 p.

APPENDIX A

ADDRESSES OF PUBLISHERS OF TESTS

American Book Company, 88 Lexington Ave., New York City; 2 North Forsyth St., Atlanta; 63 Summer St., Boston; 330 East Twenty-second St., Chicago; 300 Pike St., Cincinnati.

American Physical Education Association, Springfield, Massachusetts.

A. S. Barnes and Company, 67 West Forty-fourth St., New York City.

Bruce Publishing Company, 354 Milwaukee St., Milwaukee.

Bureau of Administrative Research, University of Cincinnati, Cincinnati.

Bureau of Educational Measurements and Standards, Kansas State Teachers College, Emporia, Kansas.

Bureau of Educational Research, University of North Carolina, Chapel Hill, North Carolina.

Bureau of Educational Research, Northern Normal and Industrial School, Aberdeen, South Dakota.

Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa.

Bureau of Publications, Teachers College, Columbia University, New York City.

Bureau of Tests and Measurements, Department of Education, University of Virginia, University, Virginia.

Bureau of Vocational Guidance, 1 Lawrence Hall, Harvard University, Cambridge, Massachusetts.

Catholic Education Press, 1326 Quincy St., Brookland Station, Washington, D. C.

Center for Psychological Service, 2024 G. St., N.W., Washington, D. C.

University of Chicago Press, 5750 Ellis Ave., Chicago.

College Book Company, Columbus, Ohio.

Columbia Graphophone Company, Woolworth Building, New York City.

Eauclaire Book and Stationery Company, Eauclaire, Wisconsin.

Educational and Personnel Publishing Company, Washington, D. C.

Ginn and Company, 15 Ashburton Place, Boston; 95 Luckie St., Atlanta; 2301 Prairie Ave., Chicago; 199 East Gay St., Columbus, Ohio; 1913 Bryan St., Dallas; 70 Fifth Ave., New York City; 45 Second St., San Francisco.

Gregg Publishing Company, 20 West Forty-seventh St., New York City; Statler Building, Boston; 2500 Prairie Ave., Chicago; Phelan Building, San Francisco; 57 Bloor Street West, Toronto.

Harlow Publishing Company, Oklahoma City, Oklahoma.

D. C. Heath and Company, 231–245 West Thirty-ninth St., New York City;

29 Pryor St., N.E., Atlanta; 285 Columbus Ave., Boston; 1815 Prairie Ave., Chicago; 1911 Bryan St., Dallas; 182 Second St., San Francisco.

Henry Holt and Company, 1 Park Ave., New York City; 6 Park St., Boston; 2626 Prairie Ave., Chicago; 149 New Montgomery St., San Francisco.

Houghton Mifflin Company, 2 Park St., Boston; 2500 Prairie Ave., Chicago; 1909 Bryan St., Dallas; 386 Fourth Ave., New York City; 612 Howard St., San Francisco.

Indiana University Book Store, Bloomington, Indiana.

Johns Hopkins Press, Baltimore, Maryland.

Joliet Township High School, Joliet, Illinois.

Kenyon Press Publishing Company, Wauwatosa, Wisconsin.

La Fayette Printing Company, La Fayette, Indiana.

Lakeland Publishing Company, 217 North Mill St., Madison, Wisconsin.

J. B. Lippincott Company, 227 South Sixth St., Philadelphia; 2244 Calumet Ave., Chicago.

Lyons and Carnahan, 221 East Twentieth St., Chicago.

The Macmillan Company, 60 Fifth Ave., New York City; 500 Spring St., N.W., Atlanta; 240 Newbury St., Boston; 2459 Prairie Ave., Chicago; Ross Ave. and Akard St., Dallas; 350 Mission St., San Francisco.

Manual Arts Press, Peoria, Illinois.

National Council of Teachers of English, 506 West Sixty-ninth St., Chicago.

National Publishing Society, Mountain Lake Park, Maryland.

Palmer Company, 120 Boylston St., Boston.

Playground and Recreation Association of America, 315 Fourth Ave., New York City.

H. C. Pryor, Kansas State Teachers College, Pittsburg, Kansas.

Public School Publishing Company, Bloomington, Illinois.

Publishing Department, Russell Sage Foundation, 130 East Twenty-second St., New York City.

Research Service Company, 7219 Beverly Boulevard, Los Angeles.

Rural Education Department, Pennsylvania State College, State College, Pennsylvania.

Safety Electric Heater Company, 761 Fourth Ave., Faribault, Minnesota.

Scott, Foresman and Company, 5 West Nineteenth St., New York City; 29 Pryor St., N.E., Atlanta; 623 South Wabash Ave., Chicago; 2013 Jackson Place, Dallas.

Smith, Hammond and Company, Atlanta.

Southern Publishing Company, Santa Fe Building, Dallas.

South-Western Publishing Company, 542 South Dearborn St., Chicago; 1–3 Third St., Cincinnati.

Stanford University Press, Stanford University, California.

D. Starch, 1374 Massachusetts Ave., Cambridge, Massachusetts.

C. H. Stoelting Company, 424 North Homan Ave., Chicago.

University Printing Company, 315 Fourteenth Ave., S.E., Minneapolis.

University Publishing Company, 1126 Q. St., Lincoln, Nebraska; 2126 Prairie Ave., Chicago; 2013 Jackson St., Dallas; 239 Fourth Ave., New York City.

C. W. Waddell, 855 North Vermont Ave., Los Angeles.
Warwick and York, 10 East Centre St., Baltimore.
John C. Winston Company, 622–33 South Wabash Ave., Chicago.
Women's Division, National Amateur Athletic Federation of America, 370 Seventh Ave., New York City.
World Book Company, Yonkers-on-Hudson, New York; 110 West Peachtree St., Atlanta; 14 Beacon St., Boston; 2126 Prairie Ave., Chicago; 1307 Pacific Ave., Dallas; Portland, Oregon; 149 New Montgomery St., San Francisco; Manila, P. I.

APPENDIX B

GENERAL BIBLIOGRAPHY

In addition to the bibliographies at the ends of the chapters dealing with special topics or fields, it seems well to include also a general bibliography on testing. The references given below include practically all books and more or less similar publications dealing with testing in general. Those devoted to a particular phase of it, such as intelligence testing, testing in a particular subject or to particular uses of tests and their results, such as for classifying pupils, prognosis, and so forth, are not included.

Buckingham, B. R. *Research for Teachers.* New York: Silver, Burdett and Company, 1926. 386 p.

Carroll, R. P. *Fundamentals in the Technique of Educational Measurements.* Syracuse, New York: R. P. Carroll, University Station, 1928 179 p.

Courtis, S. A. (Chairman). "The Measurement of Educational Products," *Seventeenth Yearbook of the National Society for the Study of Education,* Part II. Bloomington, Illinois: Public School Publishing Company, 1918. 194 p.

Doherty, Margaret and MacLatchy, Josephine (Compiled by). "Bibliography of Educational and Psychological Tests and Measurements," *U.S. Bureau of Education Bulletin,* 1923, No. 55. Washington: Government Printing Office, 1924. 233 p.

Fenton, Norman and Worcester, D. A. *An Introduction to Educational Measurements.* Boston: Ginn and Company, 1928. 149 p.

Gilliland, A. R. and Jordan, R. H. *Educational Measurements and the Classroom Teacher.* New York: The Century Co., 1924. 269 p.

Greene, H. A. and Jorgensen, A. N. *The Use and Interpretation of Educational Tests.* New York: Longmans, Green and Company, 1929. 389 p.

Gregory, C. A. *Fundamentals of Educational Measurement.* New York: D. Appleton and Company, 1922. 382 p.

Hines, H. C. *A Guide to Educational Measurements.* Boston: Houghton Mifflin Company, 1923. 270 p.

Kelley, T. L. *Interpretation of Educational Measurements.* Yonkers, New York: World Book Company, 1927. 363 p.

Levine, A. J. and Marks, Louis. *Testing Intelligence and Achievement.* New York: The Macmillan Company, 1928. 399 p.

Lincoln, E. A. *Beginnings in Educational Measurement.* Philadelphia: J. B. Lippincott Company, 1924. 151 p.

McCall, W. A. *How to Measure in Education.* New York: The Macmillan Company, 1922. 416 p.

Monroe, W. S. *An Introduction to the Theory of Educational Measurements.* Boston: Houghton Mifflin Company, 1923. 364 p.

Monroe, W. S., DeVoss, J. C., and Kelly, F. J. *Educational Tests and Measurements,* Revised and Enlarged Edition. Boston: Houghton Mifflin Company, 1924. 521 p.

Odell, C. W. "Educational Tests for Use in High Schools, Third Revision," *University of Illinois Bulletin,* Vol. 27, No. 3, Bureau of Educational Research Circular No. 53. Urbana: University of Illinois, 1929. 50 p.

————. "A Glossary of Three Hundred Terms Used in Educational Measurement and Research," *University of Illinois Bulletin,* Vol. 25, No. 28, Bureau of Educational Research Bulletin No. 40. Urbana: University of Illinois, 1928. 68 p.

Pressey, S. L. and Pressey, L. C. *Introduction to the Use of Standard Tests.* Yonkers, New York: World Book Company, 1922. 263 p.

Ruch, G. M. and Stoddard, G. D. *Tests and Measurements in High School Instruction.* Yonkers, New York: World Book Company, 1927. 381 p.

Smith, H. L. and Wright, W. W. *Tests and Measurements.* New York: Silver, Burdett and Company, 1928. 540 p.

————. "Second Revision of the Bibliography of Educational Measurements," *Bulletin of the School of Education,* Vol. 4, No. 2. Bloomington: Bureau of Coöperative Research, Indiana University, 1927. 251 p.

Starch, Daniel. *Educational Measurements.* New York: The Macmillan Company, 1916. 202 p.

Strayer, G. D., et al. "Standards and Tests for the Measurement of the Efficiency of Schools and School Systems," *Fifteenth Yearbook of the National Society for the Study of Education,* Chicago: University of Chicago Press, 1916, Part I.

Symonds, P. M. *Measurement in Secondary Education.* New York: The Macmillan Company, 1927. 588 p.

Trabue, M. R. *Measuring Results in Education.* New York: American Book Company, 1924. 492 p.

Van Wagenen, M. J. *A Teachers' Manual in the Use of the Educational Scales.* Bloomington, Illinois: Public School Publishing Company, 1928. 276 p.

Wilson, G. M. and Hoke, K. J. *How to Measure,* Revised and Enlarged. New York: The Macmillan Company; 1928. 597 p.

INDEX