A COMPARISON OF SOME METHODS OF EVALUATING OUTCOMES OF LABORATORY INSTRUCTION IN HIGH SCHOOL CHEMISTRY

by

VICTOR LENNIE CHAPMAN

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS

in the School

of

Education

We accept this thesis as conforming to the
standard required from candidates for the
degree of MASTER OF ARTS

Members of the Department of

Education .....

THE UNIVERSITY OF BRITISH COLUMBIA

OCTOBER, 1952

ABSTRACT

This study compares methods to evaluate the outcomes of laboratory instruction in high school chemistry and reports the instruments developed for that purpose.

The objectives evaluated were: the ability of students in basic laboratory skills, ability of pupils in the selection of materials, apparatus and methods; and facts that are outcomes of laboratory instruction. These three objectives were selected from some fourteen general objectives gleaned from the literature pertaining to laboratory chemistry. They were chosen as representing outcomes due solely to laboratory instruction as compared with others that may have been achieved at least in part, by the routine lessons.

The experimental method was to evaluate 72 high school students of chemistry by means of:

1. a practical test of laboratory work designed to conform with the objectives chosen referred to as the criterion test.
2. a group pencil and paper test somewhat parallel to the criterion test.
3. the laboratory notebooks of the students.
4. the teacher's estimates of student progress toward the objectives.

Three classes of chemistry were evaluated in the Spring of 1952. The teacher's estimates were prepared in February from observation of the students at work in the laboratory. The laboratory reports had been marked weekly for six months prior to the experiment and the total score on fifteen reports was taken as a measure of the notebooks to assess laboratory knowledge.

In March the criterion test was administered in two sections. Section I tested chiefly manipulations and was an individual test. Section II consisted of a series of small tests based on the course of study.

About one week later the group pencil and paper test was administered to the three classes in successive class periods. The test consisted of two parts: 1. multiple-choice items, and 2. items matching diagrams with statements.

The following statistical measures were reported for all tests: mean, standard deviation, reliability. For the criterion and pencil paper test the following were also reported: internal consistency of test items with their difficulties. The validities of the items of the pencil and paper test were also reported.

The correlations between the different tests were calculated as a means of appraising the predictive value of each. The simple regression and multiple regression equations and beta coefficients for predicting the criterion from the pencil and paper test were compared. T-scores were tabled for the pencil and paper test as well as

derived scores on the basis of a mean of 63 and a standard deviation of 13, designed so as to set 50 as the critical score to cut off 15 percent of the testees.

To compare the ability of the test to predict the upper quarter on the criterion with the lower quarter, a chi-square test of significance was applied.

The following conclusions appear to be defensible:

1. The group pencil and paper test, in predicting the criterion, was significantly superior to other methods.
2. The laboratory notebooks failed significantly to predict the outcomes being tested.
3. The teacher's estimates did not materially assist the pencil and paper test to predict the outcomes being tested.
4. The two tests possess a range of difficulty conforming to the requirements of a good test.
5. The test items having indices of validity of less than .23 contribute little to the predictive value of the pencil and paper test.
6. The pencil and paper test predicts the criterion equally well at either the upper or lower levels.

ACKNOWLEDGMENTS

This study was made possible through the cooperation of Mr.E.L.Yeo, Principal of Britannia High School, Vancouver, British Columbia. The author is further indebted to Mr.Yeo, and also indebted to Mr.C.W.Abercrombie, for reading the manuscript and for valuable suggestions.

To Mr.J.T.Young, Mr.E.H.Vollans and Mr.G.L.Phillips, the writer is grateful for their service in administering the experimental test to their classes prior to its revision, and for their forthright criticism and worthwhile suggestions. To Dr.J.R.McIntosh, for his guidance both in this study and in the preparation of the manuscript, the writer is especially indebted.

V.L.C.

CONTENTS

LIST OF TABLES

CHAPTER I

THE PROBLEM

GENERAL

For over forty years methods of laboratory instruction have been under discussion and investigation. The failure to arrive at any definite conclusion has been due, chiefly, to the conflict in the findings of the investigators. Two notable studies that failed to agree were those of Kiebler and Woody[1] and Horton.[2] The former, an earlier study, distinctly favored the demonstration method, while the latter strongly supported the individual method. The situation was further complicated when a number of schools began placing a new emphasis on certain objectives with respect to the scientific method and the scientific attitude. With the re-orientation of the objectives for secondary education, and with the new philosophy of "education for everyman's child", the secondary school population increased rapidly. In this connection there arose a demand for general science education, without detailed technical knowledge. With the increased school population, the expense of supplying laboratory equipment rose sharply.

---

1 Kiebler, E.W., and Woody, Clifford, The Individual Laboratory Versus the Demonstration Method of Teaching Physics, Journal of Educational Research, 7:50-58, January, 1923.

2 Horton, Ralph E., Measurable Outcomes of Individual Laboratory Work in High School Chemistry, (Teachers College Contribution to Education, No.303.) New York, Bureau of Publications Teachers College, 1928, p.105.

Hence any means of holding or reducing costs became urgent. Consequently, the less costly demonstration method gained favor. At the same time, some of the more important objectives of individual laboratory instruction were lost sight of. Particularly was this true where there was little or no opportunity for students to handle apparatus and reagents.

Nevertheless, laboratory work is an integral part of the training of a true scientist and since high school special science courses are generally prerequisites for this training they ought to reflect the elements of the training, even to laboratory instruction.

If desirable objectives for laboratory work can be justified, then there is an obligation to appraise the progress of students toward these objectives, in the most valid, reliable and convenient method available. At present there appear to be five methods in use. Weaknesses are evident in all these methods:

1. Marking the students laboratory notebooks

It is conceivable that a neatly written and carefully prepared book of assignments may in no way indicate the student's ability to perform an experiment, or to manipulate apparatus. It is possible that he may have plagiarized his reports from the book of a student of a previous year. It is even possible that he may have submitted the work of another student.

2. Estimating the student's laboratory proficiency

Since a person's estimate may vary from time to time, and

since different persons' estimates of the same student often disagree, they are probably highly unreliable. The variance could be due to the methods of estimating. It could also be due to the differing standards of judging as well as to the changing of standards while judging.

3. Marking the chemical products prepared in the laboratory

While this method may evaluate some of the outcomes of the laboratory, it may also lead to one of the greatest failures of traditional laboratory instruction; viz., the failure to promote growth in scientific integrity, by permitting students to submit substitutions for the products they have prepared in the laboratory.

4. Keeping attendance records

It is difficult to conceive how attendance alone can contribute to outcomes of instruction without there being some evidence of time profitably spent. However, an attendance record as a check on experiments performed would certainly have some merit when the reports were being scored.

5. Administering a practical test

Providing the test were valid and reliable it would probably be the best test of progress in laboratory work as it would be appraising either identical or related elements of the laboratory in their natural setting, the laboratory. This method appears not to be in general use, probably because it is so time-consuming.

Since the foregoing objections may be raised in connection

with the usual methods of appraisal, there seems to be need for a new approach to laboratory evaluation. In this connection a group pencil and paper test that would conform to the requirements of a good test and at the same time measure the attributes demanded by the objectives suggests itself. The advantages of such a test would seem to be:

1. It would save time

Practical tests are, as a rule, considered to be most suitable to evaluate skills of manual dexterity. They are, however, usually individual tests and as such are very time-consuming in comparison with a group test, with which as many as thirty candidates at a time may be supervised by one examiner in contrast to one candidate. Furthermore, it usually requires more time to perform a task than to select an answer to an objective test item.

2. It would be easy to administer

Printed objective group tests with directions to examiners are not difficult to administer, can be reliable, and can usually be scored by a clerical staff. On the other hand, a practical test, to be reliable, requires an experienced and capable administrator whose judgments have a minimum of variability.

3. It would be more reliable

The reliability and the validity of the scores on a test are affected by the methods of scoring, as well as by the conditions under which the test is administered. Various studies have shown that the scores of a test, when the marking is done objectively,

are more reliable than the scores when the marking requires the subjective judgment of the examiner. When the conditions under which a test is administered are subject to a high degree of control, scores are more reliable than when the conditions are subject to little control. In administering the group pencil and paper test to different classes the external conditions can be well controlled. It would be very difficult to administer the individual practical laboratory test to different groups and control such external factors as the time of the day, the mood of the examiner, and the physical conditions of the laboratory. For these reasons the group pencil and paper test would appear to be more reliable than the practical laboratory test when both tests are being administered by different examiners.

4. It could be used as a basis of promotion

Providing the group test does possess the advantages listed under headings 1, 2, and 3, then an attempt might be made by some authorities to replace current promotional practices with the better measuring instrument.

5. It would be useful to evaluate teaching

Tests that are not too lengthy and are easily scored may have some diagnostic value, particularly from the point of view of detecting gaps in the instruction of students. Teachers should welcome any device that could be used for such a purpose.

6. It could be used to investigate some phases of the learning process

It would be interesting to know what effect a thorough training in one laboratory science would have on one's ability in another laboratory science. It has been suggested that certain attitudes, such as care with delicate instruments and confidence in the use of apparatus, may be transferred from training in one science to another. It is only by investigations of these unknown educational processes that teaching can be advanced and modified. The development of tests that can give a measure of achievement in any field of endeavor has its place in assisting to discover some relationship in another field.

7. It might indicate methods of evaluation at the college level

If a pencil and paper test can be shown to correlate highly with actual performance in the laboratory at the high school level, it would point the way to a similar test for measuring the extent to which the laboratory is achieving its objectives at the advanced level.

Hendricks sums up some of the subtler advantages of such a test as follows:

> If a pencil and paper test can be developed that will have only tolerable validity, it will help to determine what our chemistry teaching program is doing. To be more specific, if we knew with some certainty just what our laboratory is doing for our students we could review our procedures with more confidence and eliminate useless parts.[1]

---

1 Hendricks, B.Clifford, "Pencil and Paper Tests in the Laboratory," Journal of Chemical Education, 22:543, November, 1945.

Mallinson[1] comes to the conclusion that there is need for reliable and valid tests for evaluating the outcomes of science teaching, other than the acquisition of factual knowledge. If the objectives of science teaching now considered of prime importance are accepted, then it would be desirable to have valid instruments to measure their attainment. This is a considered opinion after reviewing some eighty-four articles, all but nine of which were published between 1940 and 1948.

For these reasons it would seem feasible to investigate the possibility of testing some outcomes of the laboratory by means of pencil and paper tests. This, of course, will necessitate not only the determination of the objectives but also the construction of a measuring device to appraise achievement in the laboratory.

THE PROBLEM

Mention has been made of several possible methods to appraise laboratory work in high school chemistry. The problem is twofold and may be stated:

1. To prepare a valid, reliable and usable pencil and paper test pertaining to the objectives of laboratory chemistry.

2. To compare different methods of evaluating the outcomes of instruction in high school laboratory chemistry.

---

1 Mallinson, George G., "The Implications of Recent Research in Teaching of Science at the Secondary School Level," Journal of Educational Research, 43:321-42, January, 1950.

The specific methods to be employed in pursuing the investigation are:

1. An individual practical test of the objectives of laboratory instruction, to be conducted in the laboratory. This test will be called the criterion.

2. A group pencil and paper test of the same objectives as the practical test.

3. The teacher's estimates of progress in attaining the objectives of laboratory chemistry.

4. The grading of the traditional laboratory notebooks.

LIMITATIONS OF THE STUDY

The study will be limited to high school students of chemistry. Caution must, therefore, be taken in transferring any generalizations resulting from the study, to other high school sciences.

The tests, both criterion and group pencil and paper, while possessing curricular validity for students of schools in British Columbia, may well be invalid, at least in part, for students whose chemistry courses deviate from the basis of the tests.

Since the students will be tested on certain objectives of laboratory work in chemistry, the study does not presume to say how other outcomes of instruction in chemistry may be appraised.

The study will not attempt to generalize as to what is

assessed by the measures involved, except in so far as the measures involved deal with the chosen objectives.

The experimental factor will have been the method of appraising outcomes of laboratory instruction with respect to the objectives chosen.

SUMMARY

The purposes of the study are to compare methods of evaluating the outcomes of laboratory instruction in chemistry and to develop instruments for making such comparisons.

In order to investigate more fully the contributions of laboratory science, it is important to have devices for evaluation in which confidence can be placed. Experiments in the field of teaching methods require means of appraising outcomes of instruction. Until it has been found which methods are valid and reliable, little progress can be made in the methodology of laboratory science. An attempt has been made to explore several methods of appraisal with respect to the results of laboratory attainment in chemistry.

CHAPTER II

STUDIES RELATED TO THE PROBLEM

For the purpose of convenience, previous studies of testing the objectives of chemistry will be considered under the following headings: laboratory studies, tests of objectives, and standardized tests.

LABORATORY STUDIES

Of all the studies reviewed that bear on the present problem Horton's[1] is most noteworthy. In his study an attempt was made to discover if there were outcomes of laboratory instruction not tested by the typical high school chemistry examination. For the purpose of evaluating these outcomes, Horton[2] devised practical individual tests of predetermined laboratory objectives. In his tally Horton used sixteen laboratory manuals and chose 102 skills. The complete catalogue was submitted to a jury of sixteen teachers of chemistry or heads of chemistry departments in high schools. Each item was marked as: 1, habit; 2, model; or 3, to be omitted as undesirable. Of the 102

---

1 Horton, Ralph E., Measurable Outcomes of Individual Laboratory Work in High School Chemistry, (Teachers College Contribution to Education, No.303), New York: Bureau of Publications, Teachers College, 1928, p.49.

2 Ibid., p.74.

items 56.2% averaged as habits, 32.5% averaged as models and 10.4% were undesirable. From the replies to his questionnaire Horton then ranked the 55 techniques, judged to be desirable as habits.

In connection with the study Horton[1] also prepared a pencil and paper test of some fifteen diagrams and twelve statements of laboratory preparations or procedures. The student was tested on his ability to match twelve of the fifteen diagrams with the twelve statements. In the Horton study no other pencil and paper tests pertaining to laboratory achievement were used and no attempt was made to correlate the results of the practical tests with the pencil and paper test.

TESTS OF OBJECTIVES

Hendricks[2] has published some nine test items on outcomes of instruction in the chemical laboratory. The items, while not sufficient in number to form a reliable test, have published validity indices ranging from .50 to .25. Each item is prefaced by a statement of the outcome of instruction to be tested.

If a catalogue of outcomes and test items could be compiled

---

1 Horton, Ralph E., op.cit., p.74.

2 Hendricks, B.Clifford, "Pencil and Paper Tests in the Laboratory", Journal of Chemical Education, 22:543-46, November, 1945.

for laboratory sciences, then reliable and valid tests could be assembled from these items.

Numerous tests[1] have been prepared on those aspects of science teaching that are considered fundamental, and some of these are excellent. However, most of these tests measure the outcomes of science that are achieved jointly by classroom methods and the laboratory. In fact, it is conceivable that in many cases good classroom instruction without any laboratory work would show high returns on some of these tests. The following is a sample item from a test by Hendricks, Tyler and Frutchey.[2]

$$CO + 2H_2 \longrightarrow CH_3OH + \text{heat}$$

In one experiment carbon monoxide and hydrogen were heated under pressure and a catalyst to 350°C. In a second experiment, under the same conditions, but with the temperature at 1500°C. will there be any difference in the reaction and why?

(a) The reaction in the second experiment will proceed less rapidly . . . . . ( )

(b) A smaller amount of methanol will be obtained in the first experiment than in the second . . . . ( )

(c) The reaction in the second will proceed more rapidly than in the first . . . . . ( )

(d) The amount of methanol will be the same in each experiment . . . . . ( )

(e) A larger amount of methanol will be obtained in the first experiment . . . . . ( )

. . . . . .

---

1 Mallinson George C., "The Implications of Recent Research in Teaching of Science at the Secondary School Level," Journal of Educational Research, 43:321-42, January, 1950.

2 Hendricks, B.C., Tyler, R.W., and Frutchey, F.P., "Testing Ability to Apply Chemical Principles," Journal of Chemical Education, 11:611-3, November, 1934.

Check the statements that give the reasons for the answers you checked above.

(1) Temperature has no effect on rates of reaction in these experiments . . . . . ( )
(2) In this reaction an increased temperature favors the rate of reaction decomposing the product . . . . . ( )
(3) Some catalysts retard rates of chemical change . . . . . ( )

By learning the Laws of Mass Action and gaining a full understanding of them one should be able successfully to answer the questions of this type. It is conceivable that experimental evidence would help mentally to fix a principle, thus assisting in answering statement 4; but it is not necessary.

Buckingham and Lee[1] prepared a unique scheme for testing unified concepts in science. The test consisted of four parts:

(1) The student answered true-false items on the field of science to be tested.

(2) He checked those statements that he would require in order to write a theme on the field being tested.

(3) He added any significant principles he would require in his essay.

(4) He wrote the essay unifying the scientific concepts in parts 2 and 3.

---

1 Buckingham, Guy E., and Lee, Richard E., "A Technique for Testing Unified Concepts in Science," Journal of Educational Research, 30:20-27, September, 1936.

The method would seem to be worthy of consideration for the purpose of testing ability to write laboratory reports.

The open book method has been suggested by Quam,[1] and he gives a sample test. This method is useful in the classroom, but has difficulties for departmental or standardized examinations because the textbooks are not uniform. Such a test probably does measure the student's ability to use reference sources, but may also indicate his familiarity with his own text. By using diagrams or tables one might adapt the method to test ability to apply principles or to reason with scientific materials.

One of the objectives of science teaching which it is difficult to test is the ability to use the scientific method. A student may understand a general statement of the steps to be followed in the scientific method and still not be able to outline the specific steps in a logical manner or execute the procedures necessary to complete an investigation. Again, one may possess the habit of logical thinking, so necessary to apply the method, and yet lack the patience to complete an investigation. The best test of the ability to use the scientific method is to carry out an investigation even at the high school level, according to "the method". Keeslar's[2] statement of the elements of the

---

1 Quam, G.N., "Neglected Types of Examinations", Journal of Chemical Education, 17:363-5, August, 1940.

2 Keeslar, Oreon, "Elements of the Scientific Method," Science Education, 29:273-8, December 1945.

scientific method would be a good basis from which to evaluate an investigation. To get around the difficulty of the time element, however, one could use tests already developed. Such tests usually measure isolated elements of the method such as attitudes,[1] ability to apply principles,[2] or the ability to interpret experimental data.[3] It would be interesting to know how well a battery of tests of elements of the scientific method would predict the ability to apply the method in its entirety.

Webb and Beauchamp[4] devised an interesting test in laboratory resourcefulness. It was of the individual type, practical in nature, requiring the minimum of materials but considerable time to administer. The thirteen items were tabulated in order of difficulty. Laboratory resourcefulness did not find a place in the list of objectives in Appendix A, although it merits mention as an objective. It would seem that a practical test in laboratory resourcefulness could be extended and further study made in this phase of training in science. One could envisage parallel items of a pencil and paper

---

1 Ter Keunst, John, and Bugbee, Robert E., "A Test on the Scientific Method", Journal of Educational Research, 36:489-501, March, 1943.

2 Hendricks, Tyler & Frutchey, op.cit., 11:611-3.

3 Hendricks, B.Clifford, "Measuring the Ability to Interpret Experimental Data," Journal of Chemical Education, 13:62-4, February, 1936.

4 Webb, H.A., and Beauchamp, R.V., "A Test of Laboratory Resourcefulness," School Science and Mathematics, 22:259-67, March, 1922.

test that would reduce the time and labor in evaluating laboratory resourcefulness.

STANDARDIZED LABORATORY TESTS

Standardized laboratory tests have been scarce and standardized tests in chemistry have had few items pertaining to the laboratory. In most instances the same criticism is applicable, viz., these tests measure learning of a factual nature that could be achieved by studying a text or laboratory manual with diagrams of traditional laboratory experiments. One of the first of these to be published was a test by Persing[1] in which the items were related chiefly to the preparation and collection of gases.

The Stanford Aptitude Test[2] has some ingenious test items that could be adapted to testing a number of the objectives of chemistry instruction.

The Ruch-Popenoe General Science Test[3] is very factual and tests little of the other objectives of the laboratory.

The University of Chicago tests in Educational Progress in

---

1 Persing, K.M., Persing Laboratory Chemistry Test, (Form A), Bloomington, Ill., Public School Publishing Co.

2 Zyve, D.L., Stanford Scientific Aptitude Test for High School and College Students, Stanford, Cal., Stanford University Press.

3 Ruch, G.M., and Popenoe, H.F., Ruch-Popenoe General Science Test, Yonkers on Hudson, New York, World Book Co.

Biological Sciences[1] have items that are excellent for testing outcomes of the biology laboratory and are much better than the physical science counterpart for a similar purpose. This, it would be said, is in no way a condemnation of the latter test which is excellent for testing many of the general objectives of the subject.

The paucity of good standardized tests in laboratory performance makes it desirable to have studies conducted to improve the situation, not only in chemistry, but also in all laboratory sciences. Only when this is done may the revisions in our teaching methods be instituted with a background of knowledge based on experimental evidence.

SUMMARY

Of those studies reviewed in connection with testing the outcomes of laboratory instruction, Horton's[2] is the only one that considers outcomes other than those tested in a typical high school chemistry examination.

A number of excellent tests dealing with objectives in chemistry instruction have been published. Most of these tests, however,

---

1 University of Chicago, Tests in Educational Progress in Biological Sciences, (Study of Educational Progress), Chicago, University of Chicago.

2 Horton, Ralph E., Measurable Outcomes of Individual Laboratory Work in High School Chemistry, (Teachers College Contribution to Education, No.303), New York: Bureau of Publications, Teachers College, 1928, p.105.

measure factual information, or the attainment of objectives that may be achieved in part, by classroom instruction, and in part, by laboratory work.

The few standardized laboratory tests listed by publishers, and the standardized chemistry tests reviewed for this study, appear to test few, if any, of the outcomes of objectives achieved solely by laboratory chemistry.

CHAPTER III

THE PROCEDURE

OBJECTIVES

Before it was possible to proceed with the preparation of the testing devices it was necessary to have a list of acceptable objectives. For the purpose of this study, eight lists of objectives of laboratory chemistry were studied in order to choose those general objectives ranked most often.

The list of objectives for teaching of chemistry in the 46th yearbook of the National Society for the Study of Education[1] was taken as a basis. These objectives were broken down into more specific ones and in some cases reworded. To these were added any additional ones from the seven other sources. A frequency distribution was made of the listed objectives, which were then written in order of recurrence and examined to determine which were applicable to training in the laboratory.[2] From the list the following were chosen as those that should be distinctly achievable in the laboratory, as compared to others whose achievement accrues in part, at least, from daily class methods of science teaching.

---

1 National Society for the Study of Education, Science Education in American Schools, (Part I), Chicago: The Society, 1947, p.25.

2 See Appendix A, p. 74.

OBJECTIVES FOR INSTRUCTION IN THE HIGH SCHOOL CHEMISTRY LABORATORY

1. Ability to perform basic laboratory skills.
2. Ability to select appropriate materials and apparatus.
3. Ability to make accurate observations.
4. Ability to recall and use facts that are an outcome of laboratory instruction.
5. Ability to make an accurate record of observations.
6. Ability to write an acceptable piece of scientific literature or a report.

For the present study it was decided to concentrate on objectives 1, 2 and 4 of the above list. The three general objectives were separated into specific objectives. The specific laboratory skills chosen as most suitable for the present purpose was the Horton list of fifty-five basic techniques.[1] The ability to select suitable materials and apparatus could be tested to a degree in appraising the basic manipulations. The facts that are the outcome of the laboratory instruction would have to be determined prior to the experiment. In order that they be curricularly valid it was necessary to choose these objectives from the chemistry program (entitled Chemistry 91) of the students involved. Tests relating to the outcomes of the actual experiments of the chemistry course would serve to assess objectives 2 and 4.

---

1 Horton, Ralph E., Measurable Outcomes of Individual Laboratory Work in High School Chemistry, New York: Bureau of Publications, Teachers College, Columbia University, 1928, p.49.

THE CRITERION TEST

It was decided in setting up the criterion to use a revision of tests prepared by Horton[1] supplemented by a test of Chemistry 91 laboratory learning, which was prepared for this study.

Horton's test entitled "Individual Performance of Laboratory Manipulations"[2] consists of seven parts, each of the first four parts of which could be administered to a class of twenty-five students in one period of fifty minutes. Items five to seven would require about fifteen minutes per pupil. A testing procedure of this latter type would make it difficult to obtain comparable results in testing a class of twenty-five pupils and would also consume eight periods of about fifty minutes. In order to reduce the time consumed and also to cover the majority of the class in one period items five to seven of Horton's original test[3] were revised into three more balanced tests[4] of approximately the same elements.

The test of learning of Chemistry 91 from laboratory experiments was designed so that each student being tested worked simultaneously on a different test item and the group was rotated every four

---

1 Horton, R.E., op.cit., p.74.

2 Ibid., p.74.

3 Ibid., p.74.

4 See Appendix C1. p.80.

minutes. In this way ten students could do ten test items in forty minutes. By duplicating the test materials twenty pupils per period could be accommodated providing there were laboratory places available.

The score on the revision of the Horton test totals thirty-four (34) items and that of the Chemistry 91 Laboratory test totals thirty (30) items. Adding the two scores sets up a measure of the achievement of at least two phases of Chemistry 91 laboratory work. It may be argued that in combining the two tests standard scores rather than raw scores should be added. However, since the two tests are aspects of the same criterion and since the rank orders of the students in the tryout tests did not differ materially it was deemed satisfactory to add the raw scores.

All reliability formulas are based on the assumption that the greater the number of items the greater the reliability. It is advisable therefore, to lengthen tests with valid items, but not beyond the point that they become unwieldy. Some sixty items requiring at least sixty minutes of each pupil's time and requiring an estimated average of twenty minutes per pupil of the teacher's time is as much as the traffic will bear.

However, Davis[1] claims that

> so great is the importance of having a criterion variable which measures the real objective of a selection program that no effort should be spared to obtain quantitative measurements of as many elements of the real objective - the ultimate criterion - as possible, even if these measurements can be made with a reliability only slightly greater than zero.

---

1 Davis, F.B., Utilizing Human Talent, Washington, D.C., American Council on Education, 1947, p.64.

While Davis is speaking chiefly of a personnel selection program, he nevertheless indicates that his statement is applicable to tests in academic subjects; and he amplifies this point at some length. In fact, one of the chief implications of the armed services Testing Program of the United States is the necessity of validating tests and school marks against realistic criteria.

It should be pointed out again that Horton's criterion test was the result of very careful screening of objectives from numerous texts by a jury of competent chemists and teachers. The addition of items from the Chemistry 91 course of British Columbia would serve to include objectives of practical chemistry that are not solely manipulative but also interpretative.

The reliability coefficient of Horton's test[1] is given as .78 by the split-half method; the practical test for Chemistry 91 on a trial run gave a value for Rho of .84 (n = 32). The criterion, then, is a composite test that appears to fulfil the four prime considerations of validity, reliability, face validity and practicability.

THE PENCIL AND PAPER TEST

Items for this test were prepared to parallel as nearly as possible the actual items tested in the two parts of the practical tests. This was impossible in some instances, since the choices in

---

1 Horton, op.cit., p.74.

one question would undoubtedly have acted as cues for another related question. However, it does not matter too much that all items are not parallel, as the real test of the predictive value of the pencil and paper test is how well it correlates with the criterion.

After some preliminary consideration, it was decided to adhere to the multiple choice type of question in Part I, and items were prepared in this form with five choices per item. For less than five choices per item the factor of guessing is rather too high. More than five choices increases the time to administer the test, and the gain in reducing guessing is not worth the time consumed. Guessing is better handled by composing more attractive misleads. Furthermore, the difficulty of preparing six or seven choices of an attractive nature is great. It is obvious that a test item with several misleads so poor that no student chooses them becomes really a test item of only a few choices.

All items were revised in an attempt to minimize any ambiguity that appeared to exist as well as to eliminate cues. Where items had several parts, care was taken to avoid the situation where a given wrong answer in one part would affect the calculations of a later answer.

The pencil and paper test in its final form may be seen by referring to Appendix D[1] of this study. However, several typical

---

1 See page 93.

questions are cited below.

1. A typical multiple-choice item to parallel a criterion test item

Criterion test item

The student was confronted with a beaker of solution, a glass plate, a stirring rod, a vial of red litmus paper, and a vial of blue litmus paper.

Pinned to the table was the following question.

> A student has been preparing common salt by neutralization. Use the stirring rod and litmus paper to test the solution in the beaker marked '4'. Answer this question on your sheet.
>
> Should the student add a solution of (1) acid, (2) base, (3) neither, ... ( )

The students had been issued answer sheets[1] and had been instructed to follow the directions and to write the answers to the practical questions in the appropriate spaces on the answer sheets.

Group pencil and paper test to parallel the preceding item

> A student was preparing common salt by neutralization. On testing with litmus paper he found that the pink litmus paper became blue. What should he do?
>
> (1) Add a few drops of acid and test again.
> (2) Add a few drops of base and test again.
> (3) Add nothing, it is neutral.
> (4) Remove the litmus paper before evaporating.
> (5) Add a few drops of salt water to replace those used in testing . . . . . .( )

2. A typical multiple choice item not parallel to the criterion test item but later shown to have a high internal consistency and validity

---

1 See Appendix F, p. 102.

If you wished to compare the rates of reaction at two different temperatures, the most convenient temperatures to use would be: (1) 20° and 100°, (2) 10° and 90°, (3) 20° and 80°, (4) 4° and 100°, (5) 30° and 50° . . . ( )

Part II of the pencil and paper test was composed of Horton's[1] test-matching diagrams of laboratory situations with statements describing those circumstances. Little revision was attempted except to rearrange the statements in order of difficulty after the tryout.

3. Typical matching item to parallel criterion test item

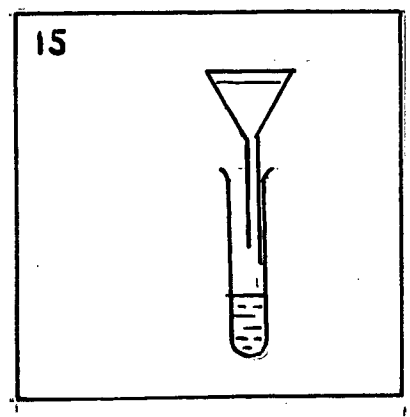
Criterion test item

Prepare a filter and filter one-third of a test tube of a liquid in bottle number I into a beaker.

The pupil's work was scored on a check sheet.[2]

Group pencil and paper test item

Apparatus to obtain quickly a suspended solid from a solution . . . . . ( )

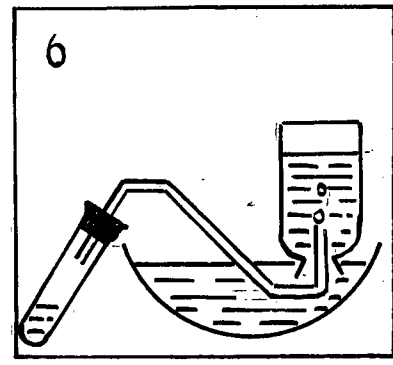


1 Horton, Ralph E., Measurable Outcomes of Individual Laboratory Work in High School Chemistry, New York: Bureau of Publications, Teacher College, Columbia University, 1928, pp. 72-3.

2 See Appendix E, p. 101.

4. Typical matching item not parallel to the criterion test but later shown to have a high internal consistency and validity

Apparatus to prepare hydrogen . . . . . ( )



Preliminary Administration

Two classes of chemistry 91 students of Britannia High School were tested in the Spring of 1951.

One-half of the pencil and paper test was administered and marked but the papers and marks were withheld. The criterion test was then administered over several weeks. Care was taken to eliminate as much as possible a leakage of test information by:

1. Testing all students of a class on one particular item at a time.
2. Testing the two classes on the same items on the same half of the school day.

Following the practical test, the remainder of the pencil and paper test was given.

In order to improve the test it was subjected to the following analysis:

1. Correlation with the criterion.
2. Item analysis.
3. (a) Validity coefficients.
   (b) Difficulty coefficients.
4. Reliability coefficients.
5. Analysis of responses.
6. Editing of items.

Correlation with the Criterion

In the preliminary tryout, by rank difference a correlation of .63 (n = 25) was shown between the test and the practical criterion test. Following the item analysis, a second correlation was computed by the same method after deleting all items from the pencil and paper test that had a validity coefficient of less than .15. Rho for this calculation was .72.

Item Analysis

For this analysis it was decided to use Thorndike's chart[1] adapted from Flanagan's abac[2]. This chart requires the top and the bottom twenty-seven percent of the papers to be analyzed so as to give the percentage of successful responses for each item in the upper and lower groups. From these two values the validity coefficient (Pearsonian r) can be read off the chart. In a study reported by

---

1 Thorndike, Robert L., Personnel Selection, New York: John Wiley and Sons, Inc., 1949, Appendix B, p. 347-351.

2 Flanagan, J.C., "General Considerations in the Selection of Test Items and a Short Method of Estimating the Product-Moment Coefficient from data at the Tails of the Distribution", Journal of Educational Psychology, 30:678, December, 1939.

Kelley[1] it is shown that the ratio of the obtained difference to its standard error is a maximum when the top and bottom group includes approximately twenty-seven percent of the population tested. Kelley states that the most satisfactory item validity index based on the upper and lower twenty-seven percent is the estimate of the coefficient of correlation between item and test obtainable from tables prepared by Flanagan.[2]

Thorndike[3] points out that,

> if the items in a test blank are examined they will be found to cover a rather narrow range in validity coefficients. An item with a validity coefficient as high as .30 usually represents an outstandingly valid item. The whole range of item validities from the most to the least may cover no more than thirty points.

Kelley suggests that an analysis for practical purposes should consist of the above method coupled with an index of difficulty based on the average of the item difficulty of the upper and lower groups.

Reliability Coefficient

The reliability coefficient of the unedited pencil and paper test using the Kuder-Richardson formula gave a coefficient of .75. The mean of the test was 18.6 and the standard deviation was 4.6.

---

1. Kelley, T.L., "The Selection of Upper and Lower Groups for the Validation of Test Items", Journal of Educational Psychology, 30:17-24, January, 1939.

2 Flanagan, op.cit., p.678.

3 Thorndike, op.cit., p.245.

The First Revision

The first revision of the pencil and paper test was made to include:

1. Those items of a validity of .15 or better, arranged in order of difficulty where possible.

2. A revision of items where a considered opinion indicated changes that would probably increase the validity by removing the ambiguity or by substituting a more suitable mislead for one that discriminates in the reverse direction. Some items were deleted and some new items were cast.

The Second Revision

In order further to improve the pencil and paper test it was administered to some sixty-four Senior Matriculation (Grade 13) students in three Vancouver High Schools, viz., King Edward, John Oliver and North Vancouver.

In order to save a year's time, the test was administered at the end of September to the students of Chemistry 100 who had taken Chemistry 91 or its equivalent the previous year. By testing in the fall it was felt that, while the results might not be as high as if the students had been tested in June, nevertheless, useful information would be at hand for the final revision of the test. The outcome of the analysis is as follows:

Number of items	64
Numbers of candidates	62
Mean score	26.45
Standard deviation	5.87 ± .73

From the item analysis it was possible to prepare a final paper of fifty items with internal consistencies of .20 or better. For the final draft each item was edited in order to replace misleads that failed to discriminate, or to recast them. The revised items were then listed in order of difficulty, re-edited and mimeographed.[1] The final draft of the paper resulted in twenty of the fifty items being related to the fifty-five basic techniques of Horton's study.[2] The remaining thirty reflected objectives of chemistry 91 laboratory learning.

THE LABORATORY NOTEBOOKS

Bulletin IX of the Department of Education of British Columbia[3] lists some thirty-one experiments which are starred in a list of fifty-eight experiments. It is intimated that the teacher should choose a suitable number of experiments including the starred list. It is possible to combine a number of the starred items into one exercise. Under the heading "Pupil Activities" it is indicated that the starred list is a minimum list and that a record of all

---

1 See Appendix D, p. 93.

2 See Appendix B, p. 95.

3 Bulletin IX Department of Education, Program of Studies for the Senior High Schools of British Columbia, Victoria, B.C.: 1939, pp. 109-115.

experiments be kept in a notebook. In a curriculum directive from the Department of Education it was indicated that twenty experiments written up would constitute an acceptable laboratory notebook provided the instructor certified the book.

For the present study the first fifteen experiments of the laboratory notebooks of the students used in the investigation were graded and the scores filed with Dr.J.R.McIntosh, School of Education, University of British Columbia, early in March, 1952, and prior to the collection of data for this study.

The Method of Scoring the Laboratory Notebooks

The experiments were each scored out of ten points with one exception where a score of twenty-two was possible. Each score was the subjective judgment of the investigator. The points kept in mind while scoring were:

1. Correct format and good English, including spelling.
2. Accuracy in procedure, materials used, formulas and equations.
3. Neatness, use of tabular outlines, legibility, and the inclusion of graphs and illustrations.
4. Originality of thought in the conclusions.

Errors were marked but no subdivision of marks was made for the above criteria. Many suggestions have been made on developing check-lists for this type of marking, one of which is that the scoring should be one point per item. However, it was felt that the average teacher marks his notebooks with less pains than perfection would re-

quire. For this study the method of the average teacher is indicated. A student's mark would be his score on the sum of the fifteen experiments.

THE TEACHER'S ESTIMATES

If teachers' estimates were valid and reliable then it would be most expeditious to use teachers' estimates in place of tests as the estimates are time-saving and labor-saving. The estimated scores, prepared by one teacher, the investigator, will be used as one means of rating laboratory performance. No generalizations can be made from estimates in this one instance, although the comparisons to be made may be of interest.

The Method of Estimating

Each student was observed during several laboratory periods unknown to him during the month of January and a subjective score given; possible 100 points. Scores ranged from 90 to 6.

A list of students and their estimated scores were filed with Dr.J.R.McIntosh in February, 1952, prior to gathering data for the investigation.

THE ASSEMBLING OF THE DATA

It appears that pupils discuss factual answers rather than methods or procedures. As a consequence it was decided to administer the criterion test prior to the pencil and paper test. By this arrangement there would probably be less discussion of what was being tested, namely, the method of doing tasks.

If the carry-over from the first test situation to the second test situation was equal in amount and in the same direction for all students, then it should have no effect on the eventual results in determining correlations. However what the transfer would be one cannot say.

Where possible, answers and scores were withheld from the student. These precautions were taken in an attempt to reduce the effect that the first testing might have on the scores of the second test.

Students were promised that a thorough discussion of all the tests would be undertaken after the completion of the testing. They were quite satisfied with the explanation inasmuch as the results would contribute to their Easter grade in Chemistry. In fact, they realized the necessity of secrecy in order not to jeopardize their grades by warning others of the test items prior to testing.

The Chemistry 91 Practical Test

The test was divided into three parts each of which was administered on successive days. Each class was divided into two parts by lot. The first part of each class was tested in three consecutive class periods on one day. The following day the remainder of each class was tested similarly. Part two of the test was administered to each half of a class in the same period; the three classes being tested in succeeding periods. Part three of the test was administered to the whole class at one sitting; thus completing

the testing of it in three successive periods. In all, this took four days to complete, but did not require all the time of every period.

To achieve this end, the investigator took the testees from their regular classes to the chemistry laboratory for the test and they returned to the regular classroom after the testing was completed. The total time that a student was absent from class would approximate fifty minutes.

The parts of the test[1] were as follows:

Part 1. Items one to six inclusive.

Part 2. Items seven to eleven.

Part 3. Items twelve to thirteen.

In administering parts one and two the test materials were set out in triplicate, that is, there were three groups of items, one group for each of four or five testees. Each student took his place at one station and performed the test. At the end of the allotted time each student moved, following chalk arrows on the floor to the next station, where he performed the next test item, and so on. In this way it was possible to accommodate fifteen students on part one or twelve students on part two of the test, at one time. An attempt was made to so place the stations that duplicate test items would be

---

1 See Appendix C, p. 79.

sufficiently far apart to prevent copying. The time was kept with a stop-watch and each student was allowed four minutes per station. At each station there was a printed sheet of instructions pinned to the table and also the required test materials.[1] Each student carried with him an answer sheet[2] on which he wrote the answers to the test. Students were instructed beforehand on the use of the answer sheet. At the completion of each part of the test the sheets were collected and scored for that part of the test. The sheets were reissued for the next part of the test at the time of testing.

Part three of the test was done by teacher demonstration. The students wrote their answers on the test blank from questions on the blackboard each of which was covered until the time of that particular test.

The Revised Horton Test

This test was administered in four parts.

Part 1.	Test 1.	Items one to eleven on the check-sheet.[3]
Part 2.	Tests 2 and 3.	Items twelve to twenty-one on the check-sheet.
Part 3.	Test 4.	Items twenty-two to twenty-eight on the check-sheet.
Part 4.	Test 4.	Items twenty-nine to thirty-four on the check-sheet.

---

1 See Appendix C, p. 79.

2 See Appendix F, p. 102.

3 See Appendix E, p. 101.

The student being tested worked behind a plywood screen at the demonstration bench and was marked by the investigator while the class proceeded with seat-work. Students averaged between three and four minutes per part of the test. In this way it was possible to test between ten and fourteen students in one class period. The students were scored directly on a check sheet using the symbol "1" for a correct response and "0" for an incorrect one. The order of testing students was by a random selection from a list of random numbers prepared by the investigator. The students were tested in the same order for each of the four parts of the test. In order not to over-penalize a student who made a blunder in part of the test, the examiner put the student right after having marked the erroneous procedure. In this test the instructions were printed and pinned to the desk and all necessary material was available.

The Pencil and Paper Test

Students were tested in three consecutive class periods by the investigator. There was no preliminary warning that a test of this nature was to be written but the students had been told that the laboratory work would be tested for the Easter reports to parents.

The papers were distributed face down after the students had been instructed as to the nature of the test. After the directions had been read and discussed the students were given exactly forty minutes to complete the test. Papers were then collected but no discussion was allowed until all classes had been tested.

SUMMARY

A list of general objectives was prepared for laboratory chemistry as a basis for evaluation of these outcomes. Two tests of chemistry laboratory attainment were devised; a practical criterion test and a somewhat parallel pencil and paper test. Every effort was made to keep these tests valid and reliable.

Seventy-two students selected from Britannia High School were rated on the basis of teacher's estimates, laboratory notebooks, the group pencil and paper test and the practical laboratory test. All possible precautions were taken to standardize the testing procedure.

CHAPTER IV

ANALYSIS OF RESULTS

METHODS OF ANALYSIS

The Reliability Coefficient

The reliabilities of all tests were computed by means of the Kuder-Richardson formula. Providing the assumptions upon which it is derived are scrupulously adhered to, this formula will give a value comparable with other methods and will avoid some of their difficulties. However, if these assumptions are not strictly followed then the results will be low. The formula is

$$r_t = \frac{(S.D.)^2 - \Sigma pq}{(\Sigma\sqrt{pq})^2 - \Sigma pq} \cdot \frac{(\Sigma\sqrt{pq})^2}{(S.D.)^2}$$

where:

p is the difficulty of each item, i.e., the percentage correct for each item.

q is 1 - p.

S.D. is the standard deviation of the test.

pq is the product for the p and the q for one item on the test.

Σ pq is the sum of all the pq's for all the items on the test.

$\sqrt{pq}$ is the square root of the product pq for one item on the test.

$\Sigma\sqrt{pq}$ is the sum of the $\sqrt{pq}$'s for all items on the test.

For the following reasons the Kuder-Richardson formula was chosen even though the optimum conditions for its use were not present.

1. The time required to administer the criterion test had been held to a minimum and to repeat the test was out of the question. Hence the test-retest procedure to determine the reliability coefficient could not be considered.

2. To avoid carry-over from one administration of the test to another, a considerable time lapse would be required. While this plan might have been arranged, there was a danger that an increase in laboratory knowledge, due to instruction in the meantime, would materially affect the scores and lower the reliability coefficient obtained for the test.

3. The test items were so dissimilar that it would have been difficult to divide the test into two comparable halves for the purpose of using the split-half method of computing reliability coefficients.

4. The tests used were not long and hence, to have reduced them to as few as twenty-five items would make them too short for the purpose of computing reliability coefficients.

The Internal Consistency of Items

The basis of internal consistency is the degree to which each item differentiates those students who are high from those who are low on the standard, i.e., the performance on the test. Each item purports to assess, in part, some simple aspect of ability. Also, the right answer for each item can be determined in advance, so it is possible to score the items on the test by a key prepared beforehand.

In validating the test it is appropriate to discover to what extent each item measures the same abilities as does the test as a whole. Nevertheless, if the test is to have breadth and scope, the indices may not be expected to be extremely high, or conversely, if the indices are very high they must be overlapping in their function as well as highly reliable. When an item index is very low, it must be either very unreliable or it measures functions quite different from the other items on the test.

So it may be said, generally speaking, that items with extremely low or negative indices are undesirable, but those of intermediate size have their place along with those that are high.

Item Analysis Indices

There are two types of situations, (1) where the performance of an item is related to a continuous measure, for example, the test score of which the item is a component; (2) where performance is being related to a dichotomy, for example, comparing the performance on an item in two groups dichotomized, say, at the median or at some level of difficulty. An adaptation of this second situation will be used in this study.

Item Indices Based on a Continuum Dichotomized for Convenience

If the testees are divided at the median the upper group may be expected to score more highly on an item than the lower group. However, if two extreme groups of, say, five percent of the total group are taken at the upper and lower level, a much greater discrimina-

tion may be expected than in the previous case. Kelley[1] has shown that the ratio of the obtained difference to the standard error of the difference is a maximum when approximately twenty-seven percent of the total testees determines the upper and lower group.

Flanagan[2] has prepared a table of product-moment correlation coefficients on the assumption that the variables responsible for item success and test score are normally distributed. One should note that in these coefficients, equal differences do not have the same significance at different levels, that is, the change from .10 to .15 is not equal to a change from .50 to .55.

According to Thorndike[3], "an item with a validity coefficient as high as .25 or .30 usually represents an outstandingly valid item."

On the basis of 72 cases the one percent level of confidence is .30 and the five percent level is .23. Hence, any item over .30 is outstanding and any below .23 should perhaps be rejected as not being significantly different from zero.

---

1 Kelley, T.L., "The Selection of Upper and Lower Groups for the Validation of Test Items", Journal of Educational Psychology, 30:17-24, January, 1949.

2 Flanagan, J.C., "General Considerations in the Selection of Test Items and a Short Method for Estimating the Product-Moment Coefficient from the Data at the Tails of the Distribution", Journal of Educational Psychology, 30:674-80, December, 1939.

3 Thorndike, R.L., Personnel Selection, New York, John Wiley and Sons, Inc., 1949, p.245.

On the basis of the present study, a comparison of the item indices computed by three methods shows them to be in agreement. The following three methods were used, of which the results of the first will be reported:

1. The upper and lower groups method according to Kelley[1] and Flanagan.[2]

2. A method of computing internal consistency utilizing the whole group and using the formula:

$$r = \frac{pq - nw}{pq}$$

where

r is the validity coefficient.

p is the proportion of students passing an item, stated as a percent.

q is 100 - p; the proportion of students failing an item, stated as a percent.

n is 100.

w is the number of students in the group q who passed the item.

3. The point biserial coefficient of correlation.

The Difficulty of Items

The difficulty of items is an important consideration in a

---

1 Kelley, op.cit., pp.17-24.

2 Flanagan, op.cit., pp.674-80.

test. Obviously, items that are passed by all testees do not discriminate, nor do items failed by all. For test construction, item difficulties require the following several considerations:

1. The highest reliability is achieved when item difficulty is at the fifty percent level, as the product of those passing and failing is at a maximum.

2. The greatest discrimination occurs when half the testees pass an item.

3. According to Adkins,

> As a general rule, the average item difficulty in a test should correspond to the average ability of the subjects; i.e., the items should be such that, on the average about half the subjects will answer correctly.[1]

If one wishes to select the top seventy percent then the difficulty should cluster around an index of .70. However, if the wish is to spread the whole group tested in rank order, then it is better to have the items of such difficulty that they range from easy to difficult with the majority at the average level of difficulty for the group.

Since the purpose of each of the tests in this investigation is to rank all the students, it would be best to have item difficulties range from easy to hard with a cluster near the .50 index level.

---

1 Adkins, D.G., Construction and Analysis of Achievement Tests, U.S. Office of Printing, Washington, D.C.: 1947, p.147.

THE DATA

The results of the tests are assembled in Appendix G, in which are listed, in order of scores on the criterion test:

1. Intelligence Quotient (IQ) as taken from student record cards. The quotients were based mainly on the Otis Self-Administering Test.
2. Raw score on the criterion (maximum - 64 points).
3. Raw score on the group pencil and paper test (maximum - 50 points).
4. Total score on the students' notebooks (maximum - 162 points).
5. The teacher's estimates (maximum - 100 points).
6. Revised Horton Test (maximum - 34 points).
7. Chemistry 91 Laboratory Test (maximum - 30 points).

Due to the absence, at various times, of different testees, scores in all data are available for seventy-two of out of some ninety participants.

THE CRITERION

On the assumption that the criterion conforms with a number of the objectives of the course in Chemistry 91, it can be said to have curricular validity. However, a study of the internal consistency and difficulty of the items, as well as the reliability of the criterion,

would permit a better judgment to be made of the ability of the test to do its appointed task.

The Reliability Coefficient (Criterion Test)

Table I based on the results in Appendix G shows the reliability of the sixty-four item criterion test to be .82.

TABLE I

SOME STATISTICAL MEASURES OF THE TESTS OF LABORATORY OUTCOMES

Measure	Range	Mean	S.D.	$S.E._m$	$S.E._{sd}$	r_t
Criterion	50-20	33.75	6.571	0.775	0.547	.823
Pencil and Paper Test	41-15	28.08	6.316	0.756	0.526	.760
Laboratory Notebooks	148-46	114.86	21.01	2.477	1.750	.117
Teacher's Estimates	86-6	52.94	18.96	2.235	1.580	.470
Revised Horton Test	24-8	17.69	3.75	0.442	0.312	.590
Chemistry 91 Lab. Test	26-7	16.22	4.20	0.495	0.350	.610

S.D. refers to the standard deviation.
$S.E._m$ refers to the standard error of the mean.
$S.E._{sd}$ refers to the standard error of the standard deviation.
r_t refers to the Kuder-Richardson reliability of the measure.

This result, therefore, appears to be sufficiently reliable to give a true picture of the status of student achievement on the

criterion. In order to raise the reliability coefficient to .90 it would be necessary to lengthen the test from 64 items to 124 items. The formula used was

$$n = \frac{r_{nn}\ (1 - r_{11})}{r_{11}\ (1 - r_{nn})}$$

where

n is the number of times the test must be lengthened to attain r_{nn}

r_{nn} is the reliability coefficient of the lengthened test.

r_{11} is the reliability coefficient of the original test.

Such an increase in the length of the test would make it too unwieldy for testing any reasonably large number of subjects.

The Internal Consistency of Items (Criterion Test)

On the basis of Flanagan's[1] table, and on the basis of the formula[2]

$$r = \frac{pq - nw}{pq}$$

internal consistencies were computed for the criterion test. A comparison is given in Table II. From this it will be seen that the indices vary from -.23 to .71. Of these six are negative and seventeen are positive but below .23.

---

1 Thorndike, R.L., Personnel Selection, New York, John Wiley and Sons, Inc., 1949, pp. 347-351.

2 See page 43.

TABLE II

CRITERION TEST

COMPARISON OF INTERNAL CONSISTENCIES BY THE METHOD INDICATED

	Flanagan	$r = \frac{pq - nw}{pq}$
Range	.71 to -.23	.79 to -.12
Median index	.32	.42
Number of items exceeding index .23	40	26
Total items	64	64

The Difficulty of Items (Criterion Test)

In computing data for Table II, item difficulties emerged routinely in the calculation of internal consistencies. The items of the criterion range in difficulty from .04 to .96 with a median of .53, a mean of .54 and one-half the items between .40 and .75. The test, therefore, is neither too difficult nor too easy, and has a desirable distribution of item difficulties.

THE PENCIL AND PAPER TEST

The Reliability of the Pencil and Paper Test

The reliability as determined by the Kuder-Richardson formula[1] gives a value of .76 which would require a test of 142 items,

---

1 See page 39.

that is, another 92 items, equivalent in every sense to the original 50 to produce a reliability of .90. The formula[1] used was

$$n = \frac{r_{nn}\,(1 - r_{11})}{r_{11}\,(1 - r_{nn})}$$

The Internal Consistency of Items (Pencil and Paper Test)

The internal consistencies of the items are compared in Table III. The three methods serve to screen out the same items in most cases.

TABLE III

PENCIL AND PAPER TEST

COMPARISON OF INTERNAL CONSISTENCIES BY THE METHODS INDICATED

	Flanagan	$r = \frac{pq - nw}{pq}$	Point Biserial
Range	.81 to .00	.60 to -.11	.55 to -.08
Median index	.37	.28	.31
Number of items over index .23	37	28	33
Total items	50	50	50

This gives us confidence in those items that are consistently good.

---

1 See page 47

The Validity Coefficients of Items (Pencil and Paper Test)

The validity coefficients were determined by the same three methods as the internal consistencies except that the individual items were compared with the total scores on the criterion test rather than with the total scores on the test itself. It will be noted by comparing Table IV with Table III that the validities of the items tend to be somewhat lower than the internal consistencies.

TABLE IV

PENCIL AND PAPER TEST

COMPARISON OF ITEM VALIDITIES BY THE METHODS INDICATED

	Flanagan	$r = \frac{pq - nw}{pq}$	Point Biserial
Range	.65 to -.33	.50 to -.22	.63 to -.19
Median index	.28	.22	.18
Number of items over index .23	30	22	20
Total items	50	50	50

The Difficulty of Items (Pencil and Paper Test)

As before, the indices of item difficulty were calculated in the preparation of item validities.

The range of difficulty is from .86 to .08 with a median of .61 and a mean of .57. The middle half of the indices ran from .44 to .71. These results compare favorably with those of the criterion.

Correlations of the Pencil and Paper Test

Table V shows the correlations between the various measures and the pencil and paper test, computed by the product-moment method. The correlation between the pencil and paper test and the criterion is .69 ± .06 from the data available. This value may be taken as the Validity Coefficient for the Pencil and Paper Test as a whole since the practical criterion test is the most sure measure of the outcomes of laboratory instruction that can be obtained.

The predictive value for a correlation coefficient of .69 can be inferred from the standard error of estimate. For the pencil and paper test predicting the criterion test the standard error of estimate is 4.55 calculated from the formula

$$S.E._t = S.D.\sqrt{1 - r^2}$$

where

$S.E._t$ is the standard error of estimate.

$S.D.$ is the standard deviation of the pencil and paper test.

r is the correlation of the pencil and paper test with the criterion test.

The value 4.55 computed from the above formula which is based on Kelley's Coefficient of Alienation, may be interpreted as follows:

When the pencil and paper test is used to predict the criterion, the chances are 68 out of a hundred that the true score would lie within ± 4.55 points of the predicted score.

Stated another way, it may be said that a correlation of .69 has an index of forecasting efficiency of 28 percent.

It should be noted that the pencil and paper test predicts the criterion to a much greater extent than it does either the manipulation of apparatus (Revised Horton Test) or the knowledge of laboratory situations (Chemistry 91 Laboratory Test). The explanation is probably twofold. Since both the criterion and the experimental test were prepared with a view to consisting of two dissimilar elements, it would be expected that the correlation between the experimental test and the criterion would be greater than between the experimental test and the two parts. The fact that the two parts of the criterion are not long would contribute to the keeping the correlations low.

TABLE V

PRODUCT-MOMENT CORRELATIONS BETWEEN THE PENCIL AND PAPER TEST AND FIVE OTHER MEASURES

Measure	Correlation
Criterion	.69 ± .06
Laboratory Notebooks	.20 ± .11
Teacher's Estimates	.67 ± .06
Revised Horton Test	.38 ± .10
Chemistry 91 Laboratory Test	.41 ± .10

The validity coefficient is affected by the reliability of the test. To reduce chance factors will increase the validity. Since lengthening the test will reduce chance factors, it will also raise the validity coefficient. It has been shown that tripling the length

of the test will raise the reliability of the test to .90.[1] If this were done the validity would rise from .69 to .75. The formula used was

$$r_{(xx)y} = \frac{r_{xy}}{\sqrt{\frac{1 - r_{xx}}{n} + r_{xx}}}$$

where:

$r_{(xx)y}$ is the validity coefficient of the lengthened test.

r_{xy} is the validity coefficient of the original test.

r_{xx} is the reliability of the original test.

n is the number of times the original test is lengthened.

It should be pointed out that since the reliabilities are probably low (due to the method of computation) the validity corrected for attenuation would probably be high. Hence, .75 may be high for the validity coefficient of this test when increased from 50 to 150 items.

THE LABORATORY NOTEBOOKS

Correlations were computed for the relation of the notebooks to the other measures in the investigation. Table VI indicates that there is a lack of relationship with the exception of the notebooks and teacher's estimates. One would surmise that the marking of a set of notebooks, weekly, would colour the teacher's judgment as to the

---

1 See page 47.

TABLE VI

CORRELATION BETWEEN THE NOTEBOOKS AND FIVE OTHER MEASURES

Measure	Correlation
Criterion	.12 ± .11
Pencil and Paper Test	.20 ± .11
Teacher's Estimates	.70 ± .06
Revised Horton Test	.06 ± .12
Chemistry 91 Laboratory Test	.22 ± .11

ability of students to do laboratory work. It is conceivable that neat, well-ordered notebooks would leave a favorable impression on the teacher that would be reflected in estimating progress. It would be well to point out the low correlation between the notebooks and the criterion. Where correlations are not substantial it indicates either, (1) marked dissimilarity, (2) unreliability, (3) coarse grouping, or, (4) non-linear relationships. In the present study the last two reasons may be dismissed, but either marked dissimilarity or the unreliability of marks assigned to the notebooks, or both, in comparison with the criterion is a possibility. Either reason would seem sufficient to deem it unworthy to use the notebook to evaluate progress of the student in laboratory work.

THE TEACHER'S ESTIMATES

The correlations of the teacher's estimates with the other measures are not high. The best correlation is with the notebooks and it has been discussed on page 54. The correlation with the criterion .47, has an index of forecasting efficiency of 12 percent as compared with one of 28 percent for the pencil and paper test. The standard error of estimate[1] of a criterion score predicted from the teacher's estimates is 5.78 which is considerably higher than one predicted by the experimental test. There is the possibility that the particulars of the pencil and paper test had so engrossed the investigator that they influenced his estimation of student achievement. This factor might account for the correlation of .67 between the test and the estimates.

TABLE VII

CORRELATIONS BETWEEN THE TEACHER'S ESTIMATES AND FIVE OTHER MEASURES

Measure	Correlation
Criterion	.47 ± .09
Pencil and Paper Test	.67 ± .06
Laboratory Notebooks	.70 ± .06
Revised Horton Test	.31 ± .10
Chemistry 91 Laboratory Test	.49 ± .09

1 See page 51.

THE REVISED HORTON TEST

For his original test, Horton[1] reported a reliability coefficient of .88 by the split half method. The revised test of 34 items as compared to 36 items of the original gave a reliability coefficient of .59 using the Kuder-Richardson formula. The original test had a median of 28.5 as compared to 17.8 for the revised test.

There may be two possible explanations for the discrepancy in the results if we assume that the conditions for administering the tests were not too different.

1. The emphasis in science teaching has changed in the last twenty-five years from the more rigorous and narrow to the less precise and general.
2. The high school student of two decades ago was more scholastically inclined than the high school student of today.

There are no marked correlations; this may be due to the unreliability of the test or to the lack of similarity between the test and the correlatives, or to the fact that the test is short, being about one-half the length of the experimental test. If we assume that the Revised Horton test of laboratory manipulations is reliable, then it would indicate that there is considerable dissimilarity between it and the Chemistry 91 test of laboratory facts and

---

1 Horton, Ralph E., Measurable Outcomes of Individual Laboratory Work in High School Chemistry, New York: Bureau of Publications, Teachers College, 1938, p.74.

associated laboratory knowledge since the correlation coefficient is .39. The results also show that the pencil and paper test is a better measure of the combined abilities of the two tests than it is of either one individually. The correlation with the criterion is .69 (Table V) with the Revised Horton is .38 (Table VIII) and with the Chemistry 91 test it is .41 (Table IX).

TABLE VIII

CORRELATIONS BETWEEN THE REVISED HORTON TEST AND FOUR OTHER MEASURES

Measure	Correlation
Pencil and Paper Test	.38 ± .10
Laboratory Notebooks	.06 ± .12
Teacher's Estimates	.31 ± .10
Chemistry 91 Laboratory Test	.39 ± .10

THE CHEMISTRY 91 LABORATORY TEST

The correlations of the Chemistry 91 Laboratory test do not run high, perhaps because of its shortness, and perhaps also because there may be a lack of relationship with the correlatives. Since checking laboratory notebooks emphasizes, in the mind of the teacher, the experiments performed, the teacher's estimates would be expected to show some correlation with the laboratory experiments, a correct assumption. ($r = .41$).

TABLE IX

CORRELATIONS OF THE CHEMISTRY 91 LABORATORY TEST AND FOUR OTHER MEASURES

Measures	Correlations
Pencil and Paper Test	.41 ± .10
Laboratory Notebooks	.22 ± .10
Teacher's Estimates	.49 ± .09
Revised Horton Test	.39 ± .10

THE MULTIPLE REGRESSION EQUATION

This investigation is concerned with deriving the best method of assessing a student's worth on the criterion. Hence, a multiple correlation was run between the criterion, on one hand, and the pencil and paper test and the teacher's estimates on the other. The resulting correlation was .6901. The correlation between the pencil and paper test and the criterion has been reported as .69. The extremely small increase in the correlation is indicative of the negligible amount the teacher's estimates contribute to predicting the criterion when combined with the group pencil and paper test.

The multiple regression equation was derived to be:

$$X_1 = .707\ X_2 + .0048\ X_3 + 13.64 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ (1)$$

where:

X_1 is the predicted criterion score.

X_2 is the actual pencil and paper test score.

X_3 is the actual estimate by the teacher of laboratory progress.

This equation shows the relative influences of X_2 and X_3 in predicting the criterion. The maximum value of the term .0048 X_3 can only be .48 which is about one-half point in 50.

THE BETA COEFFICIENTS

To get a clearer picture, the Beta coefficients were computed and compared. These standard partial regression coefficients show the relative importance of the two variables X_2 and X_3 to predict variable X_1, disregarding the differences in standard deviation.

For the variable X_2; $\text{Beta}_{12.3} = .6807$

For the variable X_3; $\text{Beta}_{13.2} = .0140$

The ratio $\dfrac{\text{Beta}_{12.3}}{\text{Beta}_{13.2}} = 48.5$, which indicates that the pencil and paper test is almost fifty times as important as the teacher's estimates in predicting the criterion. It must be reiterated that generalizations cannot be made from one case of teacher's estimates. However, the size of the ratio of the Beta coefficients may be explained by the fact that there is a high correlation between the pencil and paper test and the teacher's estimates which reduces the size of $\text{Beta}_{13.2}$ greatly, thus increasing the ratio.

THE SIMPLE REGRESSION EQUATION

The simple regression equation was computed to be:

$$Y = .717\ X + 13.61 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (2)$$

Compare this equation with equation (1) and note the similarity in the first and last terms on the right side. When the second term is deleted equation (1) becomes:

$X_1 = .707\ X_2 + 13.64$ . . . . . (3)

By the deletion of term $.0048\ X_3$ the predicted score is lowered by an amount of .48 points when X_3 is at its maximum.

STANDARD SCORES, DERIVED SCORES, AND PERCENTILES

The criterion score can be predicted from the pencil and paper test score by means of the regression equations. However, in some cases it is desirable to compare scores, and one with a possible 64 would not be suitable.

For this purpose, percentiles and standard scores are useful, although the ± character of standard scores is cumbersome. Derived scores have the advantage of being positive and of being geared to any predetermined standard deviation and mean. Two sets of derived scores have been determined:

1. based on a mean of 50 and a standard deviation of 10.
2. based on a mean of 63 and a standard deviation of 13.

The first is sometimes called a T-score and the second is the method employed by the Department of Education of British Columbia in scaling marks for departmental examinations.

After setting a critical score of 50 that would cut off the

lower 15 percent[1] in a normal distribution, a comparison was made of the predictive quality of the written test with respect to the upper and lower quarters of the distribution. From a 2 X 2 contingency table, the chi value was computed to be 1.12. Since it requires a chi value of 3.842 to be significant at the five percent level of confidence, the hypothesis that the test will predict equally well at any level has not been disproved.

The percentiles were interpolated from a graph prepared from the decile values as computed from the frequency distribution of data in Appendix G. These values are reported in Appendix J. There is no doubt that these results would be modified by taking a larger sample. It could also be argued that results of student's work from one school, under one teacher would tend to be more homogeneous than the whole high school population. Hence, a greater variance in the larger population would be expected, with the percentiles spread over a greater range and the derived scores would be compressed. An increase in the mean would lower the derived scores and a decrease in the mean would raise them.

ELIMINATION OF ITEMS WITH INTERNAL CONSISTENCIES BELOW .23

The tests were rescored after eliminating the items of low internal consistency and validity. Correlations were computed to compare the effects of the deletion. The results are reported in Table X. By a comparison with Table V it will be seen that the elimination of debatable items has had very little effect on the correlations.

---

1 See Appendix K, p. 108.

TABLE X

CORRELATIONS OF THE CRITERION AND OTHER MEASURES

AFTER DELETING INCONSISTENT ITEMS

Measure	Number of Items Deleted	Correlation with Criterion Reduced	Correlation with Pencil and Paper Test Reduced
Criterion	22		.67
Pencil and Paper Test	15	.70	
Criterion Reduced			.68

SUMMARY

For the purpose of analysis the following statistics were computed.

1. The reliabilities of the six measures.
2. The intercorrelations of the six measures.
3. The internal consistencies of items on the criterion and experimental tests.
4. The validities of items on the experimental test.
5. The difficulty of items on the criterion and experimental test.
6. The multiple regression equation for predicting the criterion from the experimental test and teacher's estimates.
7. The simple regression equation predicting the equation from the pencil and paper test.
8. Derived scores and percentiles.

9. Chi-square test of consistency of pencil and paper test with respect to predicting the upper and lower groups on the criterion.

10. Correlations of the criterion and experimental test after eliminating the inconsistent items.

CHAPTER V

SUMMARY AND CONCLUSIONS

The present investigation was undertaken to discover whether a carefully prepared, valid and reliable pencil and paper test of outcomes in laboratory instruction is as effective in measuring a student's worth in the laboratory as the traditional methods of evaluation.

The problem eventually was stated:

1. To prepare a valid, reliable and usable group pencil and paper test pertaining to the objectives of laboratory chemistry.
2. To compare different methods of evaluating the outcomes of instruction in high school laboratory chemistry.

After the objectives were chosen and limited, the study proceeded to measure student's achievement in laboratory chemistry by:

1. The traditional laboratory notebook.
2. The teacher's estimates.
3. A group pencil and paper test of the outcomes of the objectives chosen.
4. A practical test of the outcomes of the objectives chosen.

It has been indicated by related studies that traditional examinations have neglected the objectives of laboratory instruction and that these could be measured by practical individual tests.

These studies do not indicate to what extent pencil and paper tests could replace the practical type of test.

The subjects selected for the experiment were the students of Chemistry 91 in grades eleven and twelve in Britannia High School, Vancouver, British Columbia.

It was decided to run a preliminary investigation in which one class of students provided data for refining the measuring devices and techniques used.

The following year in March, the students' laboratory notebooks were graded and the teacher's estimates were prepared prior to the experiment proper. The practical laboratory test was administered in two parts:

1. The test of manipulation of apparatus called the Revised Horton test.
2. The test of practical knowledge in Chemistry 91 laboratory work called the Chemistry 91 Laboratory test.

About one week later the pencil and paper test was administered to all students of chemistry in Britannia High School. Complete results were obtained for seventy-two students.

To evaluate the effectiveness of the different measures to assess the student's worth in laboratory work, correlations were calculated between all measures. Simple and multiple regression equations predicting the score on the criterion from the experimental test and

the teacher's estimates were derived. Furthermore, reliabilities, internal item consistencies and validities were computed to evaluate tests and discover trends. Percentiles and derived scores were prepared for comparisons when further work on the problem is done.

A chi-square test of consistency of the pencil and paper test to predict the criterion was attempted on the basis of the upper and lower quarters of the criterion scores.

The results obtained were:

1. The pencil and paper test was a significantly better predictor of the criterion than any of the other measures used. ($r = .69$).
2. The inclusion of the teacher's estimates in the multiple regression equation did not significantly improve the predictive value of the simple regression equation.
3. The notebooks and teacher's estimates correlate to the extent of .70.
4. Of the measures tested the students' notebooks show the lowest correlation with the criterion, it being not significantly different from zero.
5. After the inconsistent items were deleted and the papers rescored the correlations between the criterion and the experimental test were not changed materially.
6. In comparing the degree to which the pencil and paper test will predict the upper and lower quarters of the criterion, chi was computed from a 2 X 2 contingency table to be 1.12, for which value the null hypothesis is not to be rejected.

CONCLUSIONS

The conclusions have been arranged in two divisions as they apply to the two divisions of the problem.

A. Conclusions with respect to the reliability and the validity of the pencil and paper test

1. The range and distribution of difficulties for the criterion and for the experimental test conform to the requirements for a good test.
2. About two-thirds of the items of the experimental test have internal consistencies of .23 or better, and about one-half the items have indices of validity of at least .23.
3. Since there is little change in the correlation coefficients by the deletion of items whose internal consistencies and validities are less than .23, it would indicate that these items do not contribute anything to the correlation.
4. By inspection, there appears to be some evidence that items of satisfactory validity but low internal consistency, or vice versa are reducing the correlation. Until more information is available regarding the indices, it would seem to be a wise compromise to drop only those items definitely invalid.

B. Conclusions with respect to the comparison of methods of evaluating outcomes of instruction in high school chemistry

1. Assuming the evaluation of laboratory abilities is best done by a practical test in the laboratory, this investigation,

based on the scores of seventy-two high school students, has shown that the best substitute for the time-consuming practical test is the group pencil and paper test, with respect to the objectives chosen.

2. It has further shown that the students' notebooks have failed to predict, significantly, the outcomes of these same objectives.

3. The teacher's estimates seem as successful in predicting the score on the students' notebooks as the pencil and paper test is in predicting the criterion. Since the only teacher's estimate possible was that made by the investigator himself, any generalizations regarding estimates must be very cautiously advanced. Even though the estimates were made well in advance, the investigator was not unaware of what the various factors in the testing program were to be. The estimates by the investigator might be expected, therefore, to agree more with the scores on the experimental test than would the estimates of another teacher.

4. Since the teacher's estimates correlate with the experimental test to the extent of .67 with the notebooks to the extent of .70 and yet with the criterion test to the extent of .47, it would appear that some element not present in the criterion is common to the other two measures. One hypothesis would suggest that the common element is related to the ability to write a report.

5. Both the multiple regression equation and the Beta coefficients indicate that the teacher's estimates do not materially assist

the group pencil and paper test in predicting the outcomes of the laboratory instruction. This conclusion is based on the similarity of the simple and multiple regression equations when the term $.0048X_3$ is deleted from equation (1)[1]. This is further indicated since the ratio of the Beta coefficients shows that the pencil and paper test is almost fifty times as important as the teacher's estimates in predicting the criterion. By computing a multiple correlation coefficient between the criterion and the combination of the pencil and paper test and teacher's estimates, it has been shown that a simple correlation of .69 was raised to only .6901. Such an increase is negligible, further strengthening the case for discarding teacher's estimates in this instance.

6. The relatively low correlation between the two parts of the criterion serve to support the contention that the criterion is composed of at least two dissimilar elements, viz., a test of manipulations and a test of laboratory knowledge.

SUGGESTIONS FOR FURTHER RESEARCH

1. Further research is indicated in the realm of testing the objectives of the laboratory. Investigations regarding the writing of a scientific report may vindicate the use of the laboratory notebook as a measuring device for attainment of

---

1 See page 58.

that objective of chemistry. Other objectives that might be tested are: laboratory resourcefulness, and the ability to apply the scientific method.

2. Similar investigations in the fields of physics and biology would seem to have their place in providing suitable devices for measuring the outcomes of laboratory work in those areas of science teaching.

3. The present investigation has only begun to probe the field of testing outcomes of laboratory instruction in chemistry. Since the validities of one-half the pencil and paper test items were below .23, the five percent level of confidence for these data, the test will require further revision before it can be used with much confidence. New items should be cast and the final form administered to a sufficiently large and representative cross-section of students to develop reliable norms and statistics.

4. In the development of test items it appears that items with diagrams tend to have greater validity than verbal items and it might be worthwhile to concentrate on pictorial or diagrammatic items.

5. The improvement of instruction depends in part on the ability to evaluate that instruction. When suitable tests of the outcomes of objectives become available, then will investigators of methods of instruction have tools to assess their efforts and point the way to better teaching, backed up by knowledge based on experimental evidence.

BIBLIOGRAPHY

Adkins, Dorothy G., Construction and Analysis of Achievement Tests, Washington, D.C.: U.S. Government Printing Office, 1947, p.292.

Buckingham, Guy E., and Lee, Richard E., "A Technique for Testing Unified Concepts in Science," Journal of Educational Research, 30:20-27, September, 1936.

Carmody, W.R., "Elementary Laboratory Instruction," Journal of Chemical Education, 12: 233-238, May, 1935.

Curtis, Francis, D., A Digest of Investigations in the Teaching of Science, (Textbooks in Science Education), Philadelphia: P.Blakiston & Son & Co., 1926, p.341.

Curtis, Francis D., "Milestones in the Teaching of Science," Journal of Educational Research, 44:161-178, November, 1950, pp.177.

Davis, F.B., Utilizing Human Talent, Washington, D.C.: American Council on Education, 1947, p.85.

Department of Education of British Columbia, Program of Studies for the Senior High Schools of British Columbia (Bulletin IX), Victoria, B.C.: The Department, 1937, pp. 105-119.

Edwards, Allen L., Statistical Analysis, New York: Rinehart and Co., Inc., 1946, p.360.

Flanagan, J.C., "General Considerations in the Selection of Test Items and a Short Method for Estimating the Product-moment Coefficient from the Data at the Tails of the Distribution," Journal of Educational Psychology, 30:674-680, December, 1939.

Fuller, Robert W., "Demonstration or Individual Laboratory Work for High School," Journal of Chemical Education, 13:262-264, June, 1936.

Hawkes, Herbert E., Linquist, E.F. and Mann, C.R., The Construction and Use of Achievement Examinations, Boston: Houghton Mifflin Co., 1936, p.491.

Hendricks, B.Clifford, "Measuring the Ability to Interpret Experimental Data," Journal of Chemical Education, 13:62-64, February, 1936.

Hendricks B.Clifford, "Pencil and Paper Tests in the Laboratory," Journal of Chemical Education, 22:543-546, November, 1945.

Hendricks, B.C., Tyler, R.W., and Frutchey, F.P., "Testing Ability to Apply Chemical Principles", Journal of Chemical Education, 11:611-613, November, 1934.

Horton, Ralph E., Measurable Outcomes of Individual Laboratory Work in High School Chemistry, (Teachers College Contribution to Education, No.303) New York: Bureau of Publications, Teachers College, 1928, p.105.

Hunter, George W., and Spore, Leroy, "The Objectives of Science in the Secondary Schools of U.S.," School Science and Mathematics, 43:633-47, October, 1943.

Keeslar, Oreon, "Elements of the Scientific Method," Science Education, 29:273-78, December, 1945.

Kelley, T.L., "The Selection of Upper and Lower Groups for the Validation of Test Items," Journal of Educational Psychology, 30:17-24, January, 1939.

Mallinson, George G., "The Implications of Recent Research in Teaching of Science at the Secondary School Level," Journal of Educational Research, 43:321-42, January, 1950.

The National Society for the Study of Education, Science Education in American Schools, (Part 1), Chicago, The Society, 1947, p.306.

Persing, K.M., Persing Laboratory Chemistry Test, (Form A), Public Schools Publishing Co., Bloomington, Ill.

Quam, G.N., "Neglected Types of Examinations?" Journal of Chemical Education, 17:363-5, August, 1940.

Richardson, M.W., and Kuder, G.F., "The Calculation of Test Reliability Coefficients Based on the Method of Rational Equivalence," Journal of Educational Psychology, 30:681-7, December, 1939.

Ruch, G.M., and Popenoe,H.F., Ruch Popenoe General Science Test, Yonkers-on-Hudson, N.Y.: World Book Co.

Schlesinger, H.J., "The Contributions of Laboratory Work to General Education," Journal of Chemical Education, 12:542-8, November, 1935.

Stewart, A.W., "Measuring Ability to Apply Principles," School Science and Mathematics, 35:695-9, October, 1935.

Ter Keunst, John and Bugbee, Robert E., "A Test on the Scientific Method," Journal of Educational Research, 36:489-501, March, 1943.

Thorndike, Robert L., Personnel Selection, New York: John Wiley & Sons, Inc., 1949, p.358.

University of Chicago, Tests in Educational Progress in Biological Sciences, (Study of Educational Progress), Chicago, University of Chicago.

Webb, H.A., and Beauchamp, R.V., "Test of Laboratory Resourcefulness," School Science and Mathematics, 22:259-67, March, 1922.

Wise, Harold E., "A Determination of the Relative Importance of Principles of Physical Science for General Education (1 and 2)," Science Education, 25:371-79, December, 1941 and 26:8-12, January, 1942.

Wrinkle, F.B., Improving Marking and Reporting Practices in Elementary and Secondary Schools, New York: Rinehart, 1947, p.120.

Zyve, D.L., Stanford Scientific Aptitude Test for High School and College Students, Stanford, Cal.: Stanford University Press.

APPENDIX A

OBJECTIVES

This list of fourteen objectives has been derived from eight sources and has been ranked in order of frequency.

		Rank
1.	Ability to make conclusions from observations.	1
2.	Ability in basic laboratory skills.	2
3.	Ability in the selection of materials and apparatus.	3
4.	Understanding of the scientific method.	3
5.	The student is developing an interest in science.	3
6.	Ability to make accurate observations.	4
7.	Ability to make an accurate record of observations.	4
8.	Understanding of principles.	4
9.	Ability to apply principles.	4
10.	Facts that are an outcome of laboratory instruction.	4
11.	Ability to write an acceptable piece of scientific literature or a report.	4
12.	Develop habits of accuracy.	4
13.	Development of attitudes.	4
14.	Appreciation of science.	4

APPENDIX B

APPROVED LIST OF LABORATORY TECHNIQUES RANKED ACCORDING TO IMPORTANCE[1]

1. Twist or screw a stopper into a tube.
2. Twist or screw a glass tube into a rubber stopper.
3. Smooth the ends of freshly cut glass tubing. (fire-polishing).
4. Always pour concentrated sulfuric acid into water - never water into concentrated acid.
5. Smell gases by fanning toward the nose - never inhaling.
6. Wash all glassware when through using.
7. Turn the water faucet off when through using.
8. Avoid pointing the mouth of the test tube containing a reaction at anyone's face.
9. Always replace reagent bottle in exact place where found immediately after using.
10. Throw all solidwaste in waste jars - not in sink.
11. Flush sink after pouring in acid.
12. Be able to cut a glass tube at any point by making a scratch with a file and then breaking with pressure.
13. Wash the table top after each experiment.
14. Avoid 'sucking back' of a delivery tube by disconnecting, or by taking the end from the water, as soon as heating is completed.

---

1 Horton, Ralph E., Measurable Outcomes of Individual Laboratory Work in High School Chemistry, New York: Bureau of Publications, Teachers College, 1928, p.49.

15. In filtering, keep the liquid below the edge of the filter paper.

16. Use the tip of the bunsen flame - not the base - when applying heat.

17. Use a flame spreader when heating glass tubing to be bent.

18. Clamp a test tube firmly but without pressure.

19. Fold a filter paper to form a smooth cone to fit a funnel.

20. Take a stopper from a bottle by turning the palm upward and holding the stopper between the fingers.

21. Hold the stopper in the hand until through using the bottle, then replace it in the bottle.

22. When washing the table, squeeze the sponge and take up excess water.

23. Dry glass vessels on the outside before heating them.

24. In heating a glass vessel move the heat around - do not heat in one place.

25. Begin to heat any vessel of glass gradually.

26. Use a wire gauze or asbestos beneath beakers and flasks when heating them.

27. Be able to adjust a ringstand clamp to any height or any angle.

28. Wet a filter paper before using it for filtering.

29. In evaporating to dryness, remove the flame before the last bit of water disappears.

30. In using a thistle tube in a generator, be sure that the lower end is below the surface of the liquid in the generator.

31. Put powders on creased papers and pour them into small mouthed bottles.

32. Without admitting air, be able to invert a bottle of water with a glass plate over the mouth beneath the water in a trough.

33. Insert the delivery tube beneath an inverted bottle of water in a trough without admitting air.

34. Set up bottles of gas, upright or inverted as determined by the weight.

35. When necessary use a pestle and mortar to pulverize coarse materials.

36. When about to light a bunsen burner, light the match before turning on the gas.

37. For ordinary use, turn the flame down to about four inches.

38. Keep the flame down below the level of the liquid in a vessel which is being heated.

39. Wet a rubber stopper when connecting it to glass and wet a glass tube when inserting it into rubber tubing.

40. Slide solids into test tube with the tube in an oblique position, to avoid breaking the tube.

41. When a crucible is to be heated select a pipestem triangle for its support on the ringstand.

42. When a dry gas, lighter than air but soluble in water is to be collected, collect it in an inverted bottle by the displacement of air.

43. When a dry gas, heavier than air, but soluble in water, is collected displace air from an upright bottle.

44. Wash and save zinc after using a hydrogen generator.

45. To correct the striking back of a bunsen burner, extinguish the flame and relight.

46. When heating a solid in a test tube, hold the tube in an almost horizontal position with the mouth slightly lower than the closed end.

47. When a funnel is to be set on the table, stand it with the mouth down.

48. Be able to make a smooth, rounded, right angle bend from a straight glass tube.

49. Test the force of water before putting a vessel beneath the faucet.

50. Be able to estimate, approximately five grams, by reference to the weight of a nickel coin.

51. In weighing, use the right hand pan for weights, placing object to be weighed on the left.

52. Read a centigrade thermometer to 0.5 of a degree.

53. Rotate a bottle when pouring powders from it.

54. Devise a condenser by surrounding a test tube with cold water in a beaker or pan.

55. Touch the sides of the receiving vessel with the end of a funnel when making filtration.

APPENDIX C

1. THE REVISED HORTON TEST

2. THE CHEMISTRY 91 LABORATORY TEST

1. REVISED HORTON TEST

Prepare a filter and filter one-third of a test tube of a liquid in bottle number 1 into a beaker.

REQUIREMENTS

1. A shelf of reagents including one marked '1'.
2. A filter stand.
3. A funnel.
4. A pack of test tubes.
5. A sink and tap.
6. A box of filter paper.

2.

Light a bunsen burner; adjust the flame for use. Correct the flame that has struck back.

REQUIREMENTS

1. A bunsen burner connected to the gascock.
2. A box of matches.

3.

Half fill a test tube with water; clamp it to the ring stand and heat it to boiling.

REQUIREMENTS

1. A ring stand and clamp.
2. A bunsen burner.
3. A rack of test tubes.
4. A box of matches.

4

Take about five grams of powder from each of the bottles '1' and '2'. Mix the powders and place in a test tube. After you have finished set it up to generate a gas by heating the mixture.

REQUIREMENTS

1. A bottle of powder marked '1'.
2. A bottle of powder marked '2'.
3. A pad of paper.
4. A rack of test tubes.
5. A piece of rubber hose.
6. A rubber stopper with glass tube inserted.
7. A spatula.
8. A pestle and mortar.
9. A beaker.

5

Set up a jar to collect hydrogen in the usual way.

Show how you would set a jar of hydrogen on the table where it is to remain for several hours.

REQUIREMENTS

1. Two gas bottles.
2. Two glass plates.
3. A pneumatic trough.
4. A $\frac{1}{4}$" rubber tube 24" long.
5. Water tap.
6. A sink.

2. CHEMISTRY 91 LABORATORY TEST

1.

The three solutions marked 1, 2, and 3 may contain iodine. Test a few c.c.'s of each solution with hypo (sodium thiosulfate) and state which contains iodine.

1. The bottle marked . . . . . contains iodine. ( )

REQUIREMENTS

Three solutions: 1. Ferric chloride.
2. Potassium dichromate.
3. Iodine and Potassium iodide solution.

Test solution: Hypo (sodium thiosulfate) solution.

2.

A student has been preparing common salt by neutralization. Use the stirring rod and litmus paper to test the solution in the beaker marked '4'.

Answer these questions on your sheet.

2. Should the student add a solution of (1) acid, (2) base, (3) neither? ( )

3. What acid or base should he use? If none, write 'nil' in the blank.( )

REQUIREMENTS

1. Slightly basic salt solution.
2. Red litmus paper.
3. Blue litmus paper.
4. A glass plate.
5. A stirring rod.

3.

DO NOT TOUCH THE BURETTE!

Before titration the burette was filled with base to the zero mark. The investigator used the pipette for the acid and completed the titration. The base is 0.20N.

4. Has the end point (1) been reached?
(2) been overrun?
(3) not been reached?
(4) been neutralized? . . . . . ( )

5. What volume of base has been used? . . . . . ( )

6. Assuming neutralization to be complete at 20.0 cc's, calculate the normality of the acid . . . . . (N.)

REQUIREMENTS

1. A burette filled to 15.3 c.c.
2. A 10 ml. pipette.
3. A beaker containing 25 c.c.'s of solution colored red with phenolphthalein.

4.

Smell each of these solutions as a preliminary test and then verify it using the reagents in front of you. If any gas is not present write "nil" in the parentheses.

7. Which solution contains hydrogen sulfide? . . . . . ( )

8. Which solution contains sulfur dioxide? . . . . . ( )

9. Which solution contains carbon dioxide? . . . . . ( )

REQUIREMENTS

1. A solution of sulfur dioxide marked "4".
2. A solution of hydrogen sulfide marked "5".
3. A solution of carbon dioxide marked "6".
4. A solution of limewater reagent (calcium hydroxide).
5. Lead acetate paper.
6. A DILUTE SOLUTION of potassium permanganate labelled 'red dye'.

5

The jars marked '7', '8' and '9' contain one each of the following: gypsum ($CaSO_4.2H_2O$), common salt (NaCl) and potassium nitrate (KNO_3). By dissolving a small portion of each in water discover which sample is:

10. most soluble in cold water . . . . . ( )

11. second most soluble in water . . . . . ( )

12. least soluble in water . . . . . ( )

REQUIREMENTS

1. A rack of test tubes.
2. A spatula.
3. A jar of sodium chloride labelled '7'.
4. A jar of gypsum labelled '8'.
5. A jar of potassium nitrate labelled '9'.
6. A pad of paper. 4" X 4".

6

DO NOT TOUCH THE BALANCE OR RAISE THE PANS!

You may handle the weights with forceps. Return the weights to the pan when you are finished.

The crucible and contents have been weighed.

13. Show how you would calculate the weight.

14. What weight has the crucible and contents? . . . . . (gm.)

REQUIREMENTS

1. A balance with (a) a crucible of salt on the left hand pan.

(b) the following weights on the right hand pan: 10, 2, and 1 grams; 500, 200, 50 and 5 milligrams.

7

One flask contains lead chloride precipitated and the other contains silver chloride precipitated. Shake each flask well and pour about 5 c.c. of the suspension into two separate test tubes.

15. Heat each test tube in turn and decide which flask contains lead chloride . . . . . ( )

DO NOT EXTINGUISH THE BURNER!

REQUIREMENTS

1. A flask of lead chloride precipitated.
2. A flask of silver chloride precipitated.
3. A burner.
4. A rack of test tubes.
5. A test tube clamp.

8

Each of the three bottles marked '12', '13', and '14' contains one of the following salts in solution; Sodium chloride, sodium bromide, and sodium iodide.

Using the chlorine water, bromine water, and benzene, test a small sample of each solution to determine:

16. Which bottle contains the iodide? . . . . . ( )

17. Which bottle contains the bromide? . . . . . ( )

18. Which bottle contains the chloride? . . . . . ( )

REQUIREMENTS

1. A solution of sodium chloride marked '14'.
2. A solution of sodium bromide marked '13'.
3. A solution of sodium iodide marked '12'.
4. A flask of chlorine water.
5. A flask of bromine water.
6. A bottle of benzene.
7. A rack of test tubes.

9

The unknown solution in bottle '15' may contain silver ions and barium ions. Test for the presence of each ion using about a 5 c.c. sample for each.

19. Does the sample contain silver ions? . . . . . ( )

20. Does the sample contain barium ions? . . . . . ( )

REQUIREMENTS

1. A solution of Silver nitrate marked '15'.
2. Hydrochloric acid reagent.
3. Ammonium hydroxide reagent.
4. Sulfuric acid reagent.
5. A rack of test tubes.

10

DO NOT CONTAMINATE THE SOLUTIONS BY CHANGING THE WIRES

Test each of the solutions '16', '17' and '18', to determine by a flame test which solution contains:

21. a barium salt . . . . . ( )

22. a sodium salt . . . . . ( )

If a solution is absent write 'nil' in the blank.

REQUIREMENTS

1. A flask of concentrated sodium chloride marked '16' and containing a flame test wire.
2. A flask of concentrated barium chloride marked '17' and containing a flame test wire.
3. A flask of concentrated calcium chloride marked '18' and containing a flame test wire.
4. A lighted burner.

11

In the rack are five precipitates of metallic sulfides. By their colours choose:

23. copper sulfide . . . . . ( )

24. cadmium sulfide . . . . . ( )

25. antimony sulfide . . . . . ( )

REQUIREMENTS

A rack of test tubes containing:

(1) zinc sulfide precipitated.

(2) antimony sulfide precipitated.

(3) manganous sulfide precipitated.

(4) cadmium sulfide precipitated.

(5) copper sulfide precipitated.

12

26. What term is best applied to the solution?
(1) superheated. (2) supersaturated. (3) oversaturated. . ( )
(4) superconcentrated. (5) hydrated.

27. The process of solidification is called: (1) crystallization.
(2) precipitation. (3) consolidation. (4) coagulation.
(5) petrifaction. . . . . ( )

REQUIREMENTS

1. A flask of supersaturated hypo (sodium thiosulfate)

2. A crystal of hypo.

13

28 & 29 What test is being performed? (. . . . . test for . . . .)

30. Was the unknown present? . . . . . (. . . .)

REQUIREMENTS

1. A solution of sodium nitrate.
2. Concentrated sulfuric acid.
3. A freshly prepared solution of ferrous chloride.

The teacher performs test 12 by adding a crystal of hypo to the supersaturated solution and showing the pupils the crystallization.

Test 13 is performed by the teacher illustrating the brown ring test for nitrates.

APPENDIX D

THE PENCIL AND PAPER TEST

CHEMISTRY 21

NAME ..... DATE ... .

Laboratory Examination.

SCHOOL ..... .

You are being tested on your knowledge of;(1) laboratory procedures you have learned in chemistry, and (2) experiments you have learned, observed or performed.

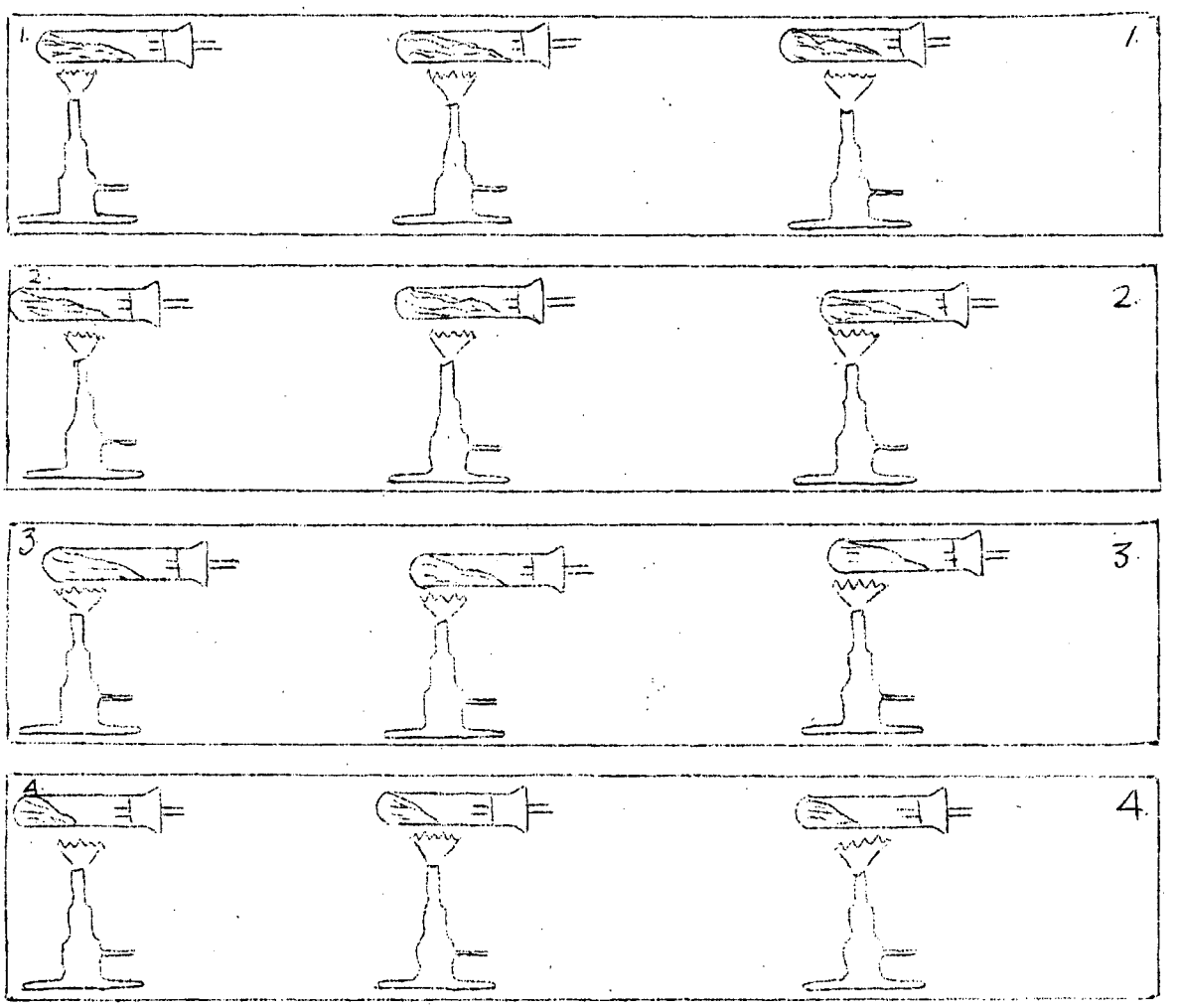
DIRECTIONS Read each question carefully and place the number of the best answer in the space provided at the right of each question.

EXAMPLE: About five grams of salt should be: (1) one-quarter teaspoonful. (2) one teaspoonful. (3) one and one-half teaspoonfulls. (4) two teaspoonfulls. (5) five teaspoonfulls. (2

A (2) is placed in the parentheses because it is the best answer.

1. When a chemist is identifying a gas by smell he should: (1) have his antidotes for poison on the bench beside him. (2) sniff it gently first and only deeply if it is not nauseating or irritating. (3) waft it gently toward him a nd sniff cautiously. (4) hold a damp cloth near his nose in order to reduce the concentration of the gas. (5) stand by an open window in case the gas is smelly. (

2. Which block of diagrams shows the correct sequence for heating a solid mixture in a test tube to produce a gas? (



3. In finding the percent of water of crystallization in a salt by heating the hydrate to the anhydride, how often should you alternately heat it and weigh it? (1) Until the calculated amount of water has been driven off. (2) Just once is enough. (3) Twice for accuracy. (4) Until the last weight is unchanged from the previous one. (5) As often as class time permits. ( )

4. A student was determining the combining weight of magnesium and found it to be 12.0 grams. The true combining weight is 12.16 grams. He made a calculation $(.16 \times 100\%)$. What was he attempting to calculate? $(\overline{12.16}$
(1) Percent yield. (2) Percent deviation. (3) Percent error. (4) Average percent. (5) Percent correct. ( )

5. Into a clear solution a small crystal was dropped. The Solution immediately solidified and became warm. What term best applies to the solution? (1) Superheated. (2) Supersaturated. (3) Oversaturated. (4) Superconcentrated. (5) Hydrated.

6. A student was preparing common salt by neutralization. On testing with litmus he found the pink litmus became blue. What should he do? (1) Add a few drops of acid and test again. (2) Add a few drops of base and test again. (3) Add nothing, it is neutral. (4) Remove the litmus paper before evaporating. (5) Add a few drops of salt water to replace those used in testing. ( )

7. After you have prepared hydrogen with zinc and HCl and are cleaning up, which step is most important? (1) Throw the unused zinc in the waste jar and pour the acid down the sink. (2) Save the acid solution and return it to the HCl winchester. (3) Burn all the hydrogen left over and so prevent an explosion. (4) Wash the acid down the sink with plenty of water. (5) Put both the acid and zinc in the waste jar. ( )

8. When hydrochloric acid is being poured from a reagent bottle, the chemist should: (1) lay the stopper on the table. (2) lay the stopper on a clean piece of glass. (3) withdraw the stopper between the fingers of the right hand with the palm facing down. (4) Withdraw the stopper between the fingers of his right hand with the palm facing up. (5) place the stopper in the rack provided. ( )

9. In lighting a bunsen burner the first thing to do is: (1) turn the gas on strong before lighting the match. (2) turn the gas on weak before lighting the match. (3) light the match before turning on the gas. (4) open the the air valve at the base of the burner before turning on the gas. (5) light the gas before closing the air valve at the base of the burner.

10. If you accidentally spilled a little spot of sulfuric acid on your coat, you should: (1) put it near the radiator so the acid will evaporate quickly. (2) put your coat in water immediately. (3) sponge the area affected with dilute ammonium hydroxide and water. (4) pour a dilute sodium hydroxide solution on the affected part. (5) sponge with water and let it dry. ( )

11. Which of the following grades is not found on labels in the laboratory storeroom? (1) C.P. (2) U.S.P. (3) Tech. (4)S.Q. (5) meets A.C.S. standards. (

12. A group of students were doing an experiment involving the differences in several readings of temperature. They decided to let one boy do all the readings and chose him by lot. Their reason for having one boy read the thermometer was: (1) if the results were poor they would know whom to blame. (2) that any errors in one person's readings would most likely be consistent and cancel out. (3) that by choosing him by lot they would not likely get the poorest person to read the thermometer. (4) too much time would be spent in arguing if more than one person read the thermometer. (5) it would fit into a plan to divide up the work in doing the experiment.

13. A student was confronted with water solutions of the following gases: (1) carbon dioxide, (2) hydrogen sulfide, (3) nitrogen, (4) oxygen, and (5) sulfur dioxide. He smelled them and chose one that smelled like low-tide. He tested it with lead acetate paper. The result was dark coloration. the gas was (

14. The solution of gas (listed in question 13) that irritated his nostrils and bleached a red dye colorless was (

15. The third solution (listed in question 13) tested had no odour but gave a white precipitate with calcium hydroxide solution. The dissolved gas was (

16. If the gas flame of a bunsen burner strikes back (i.e. burns at the base of the burner) one should: (1) turn it off and relight. (2) turn it off and get another burner.(3) close the air valve at the base of the burner and it will be corrected. (4) call the instructor and have him relight it. (5) reduce the gas pressure at the stopcock. (

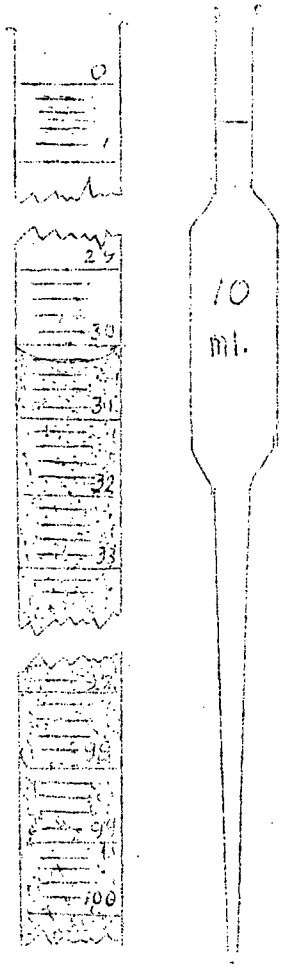
17. A student mixed some fertilizer and lime to test for ammonia. The resulting gas smelled like ammonia but did not affect either red or blue litmus paper. His most probable error was in: (1) identifying the gas by smell. (2) using old litmus paper. (3) not wetting the litmus paper. (4) using the wrong indicator. (5) using the wrong chemicals. (

18. Into a clear solution a small crystal was dropped. The solution immediately solidified and became warm. The process of solidification is best called: (1) crystallization. (2) precipitation. (3) consolidation. (4) coagulation. (5) petrifaction.

19. In making a test for an unknown acid radical the student added five cubic centimeters of freshly prepared ferrous sulfate solution to an equal volume of the unknown. He then carefully poured concentrated sulfuric acid down the inside of the test tube containing the mixture just prepared. The test performed was to test the presence of: (1) sulfate. (2) chlorate. (3) phosphate. (4) chloride. (5) nitrate radical. (

20. The name of the test described in question 19 is the: (1) sulfate test. (2) molybdate test. (3) reduced iron test. (4) oxidized iron test. (5) brown ring test. (

21. The preparation of dilute sulfuric acid from concentrated in the laboratory is a slow process because: (1) the acid is not very soluble and so takes some time to dissolve. (2) the sudden heat generated would break any common glass vessel unless it is mixed slowly. (3) The acid vaporizes and so must be kept covered. (4) sulfuric acid is oily and so it is difficult to mix it with water. (5) if the acid gets too hot it will dissolve the glass container. (

22. 

Before titration the 100 c.c. burette was filled to the zero mark. After one titration the level of the base appeared as in the diagram. The acid was delivered from the pipette shown. The base was 0.20 N. Titration was continued until the phenolphthalein indicator became a deep red. The volume of base used was: (1) 30.1c.c. (2) 31.0c.c. (3) 32.2c.c. (4) 30.2c.c. (5) 30.22c.c (

23. The end point is said to have been: (1) reached. (2) overrun. (3) achieved. (4) not reached. (5) confirmed. (

24. Assuming the burette to read 12.0 c.c. then the normality of the acid would be: (1) 0.24 N. (2) 0.4 N. (3) 0.06 N. (4) 0.60 N. (5) 0.167 N. (

25. If the experimenter wished to repeat the experiment he should take a fresh sample of acid and indicator and then: (1) use a funnel to fill the burette. (2) proceed with his titration to 62 c.c. (3) proceed with his titration until the end-point is reached. (4) drain the burette to an even volume (e.g. 40c.c.) before proceeding with the titration. (5) empty the burette and wash before filling with 0.20 N. base. (

26. How is the folded filter paper held in position in the funnel before filtering is commenced? (1) Use one hand to hold the paper and pour the liquid from the vessel with the other hand. (2) The cohesion between the dry paper and the glass will keep it in position. (3) The adhesion between the dry paper and the glass will hold it in position. (4) Wet the filter paper with your solvent after inserting it in the funnel. (5) Wet the filter paper with your solution after it is in position in the funnel. (

27. The following colours are produced by the vapours of different metals, (1) brick red, (2) light green, (3) yellow, (4) crimson red, (5) violet, and (6) blue green.
Barium would produce what colour? ( )

28. Using the colours stated in question 27, write the number of the colour produced by sodium vapour. (

29. You have three unknowns which contain (1) a chloride, (2) a bromid and (3) an iodide in solution. In order to test and identify each you would add: (1) chlorine water. (2) bromine water. (3) carbon disulfide. (4) chlorine water and then carbon disulfide. (5) bromine water and then carbon disulfide. (6) either chlorine water or bromine water and then carbon disulfide. (7) none of the methods stated above. You can only determine it by eliminating the other two halides.
Which of the above statements is the best explanation of determining: the bromide? (

30. the iodide? (

31. the chloride? (

32. You have five flasks containing yellow solutions. They are (1) impure hydrochloric acid, (2) colloidal arsenic trisulfide, (3) methyl orange indicator, (4) dilute ferric chloride solution, and (5) dilute potassium chromate solution. Which of the above will be precipitated by adding a few c.c.'s of dilute ammonium hydroxide? (

33. Which of the solutions in question 32 would be precipitated by adding a few c.c.'s of hydrochloric acid? (

34. If you wished to compare the rates of reaction at two different temperatures, the most convenient temperatures to use would be: (1) 20°C. and 100°C. (2) 10°C. and 90°C. (3) 20°C. and 80°C. (4) 4°C. and 100°C. (5) 30°C. and 50°C. (

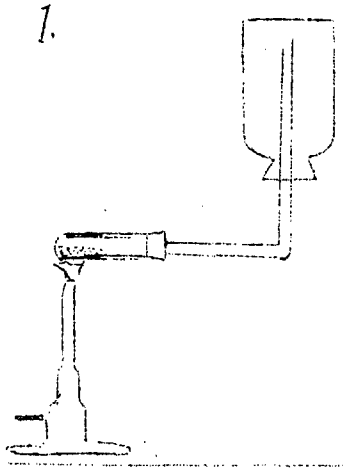
35. A sample of baking powder undergoing analysis produced the following tests: (1) the filtrate tested for sulfate. (2) the filtrate tested for phosphate. (3) No ammonium salts were in the filtrate. Which two of the following substances were definitely present in the baking powder? (1) combined calcium. (2) molybdates. (3) combined aluminum. (4) tartarates. (5) yeast. (6) Ammonium bicarbonate. (

36. (

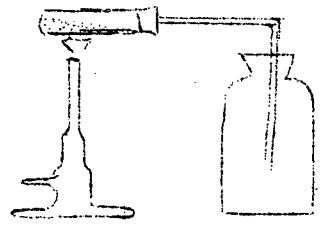
37. In the procedure of lighting a bunsen burner one should: (1) open the air valve at the base before turning on the gas. (2) turn the gas on weak until it is lighted. (3) hold the lighted match close to the burner. (4) turn the gas on strong until it is lighted. (5) test the gaspressure before attaching the burner. (

38. If you were using a 200 c.c. graduate with 10 c.c. graduations and measured out 150 c.c. of water, to which was added 120 c.c. of alcohol, what percent of the whole mixture was alcohol? Choose the answer that you can be most sure of. (1) 44%. (2) 40%. (3) 44.4%. (4) 44.44%. (5) 44.444%. (

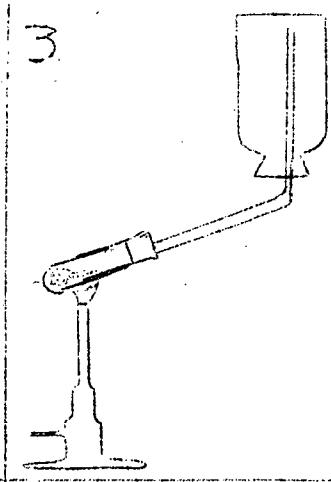
1.



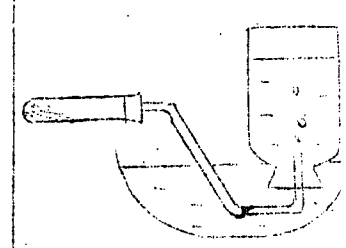
2.



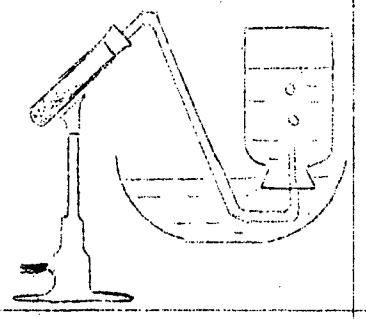
3.



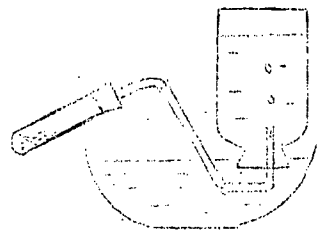
4



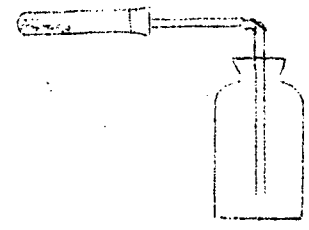
5



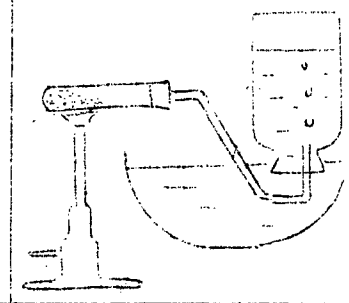
6



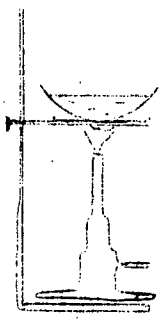
7



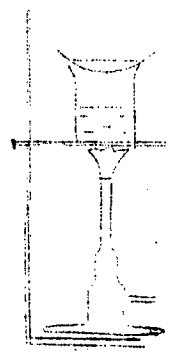
8



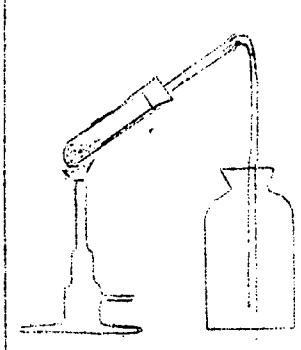
9



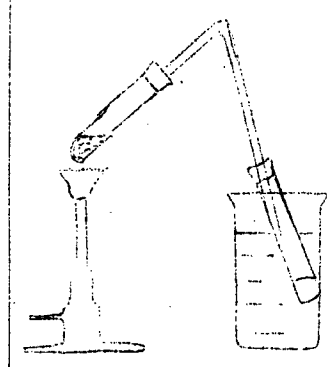
10.



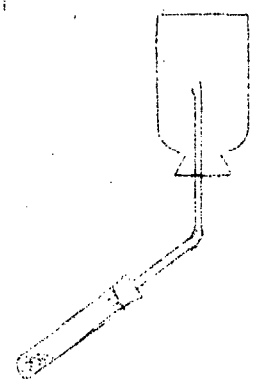
11.



12



14



15

PART II

Select from the sketches of the apparatus on the opposite page, the apparatus best designed to do the task required in each of the following cases. Write the number of the apparatus in the parentheses provided at the right.

1. Apparatus to obtain quickly a dissolved solid from solution. ( )

2. Apparatus to prepare a gas heavier than air, soluble in water and made from heating a liquid and a solid. ( )

3. Apparatus to obtain quickly a suspended solid from solution. ( )

4. Apparatus to distil water. ( )

5. Apparatus used to prepare a gas heavier than air, soluble in water and made by heating two solids. ( )

6. Apparatus used to make a gas lighter than air, soluble in water, and formed by the action of a liquid on a solid without heating.( )

7. Apparatus used to make crystals of a solid from a solution of the solid. ( )

8. Apparatus to prepare a gas lighter than air, soluble in water and made by heating two solids. ( )

9. Apparatus to prepare oxygen. ( )

10. Apparatus to prepare hydrogen chloride gas. ( )

11. Apparatus to prepare hydrogen. ( )

12. Apparatus to prepare chlorine. ( )

APPENDIX E

CHECK SHEET

for scoring

THE REVISED HORTON TEST

	Olsen C.	Jarvis A.	Brisseua G.	Goodman M.	Wong C.	Scrimgeour G.
1. Folds paper properly.						
2. Inserts paper in funnel correctly.						
3. Wets paper.						
4. Pours liquid not above paper.						
5. Touches funnel to edge of beaker.						
6. Takes stopper between fingers palm up.						
7. Keeps stopper in hand while pouring.						
8. Keeps bottle in hand until through.						
9. Replaces stopper and bottle to right place.						
10. Hold test tube obliquely.						
11. Catches last drop on edge of test tube.						
12. Lights match before turning on the gas.						
13. Turns gas on strong at first.						
14. Holds match high.						
15. Closes air inlet before lighting.						
16. Turns flame down to four inches.						
17. Puts paper in jaws of clamp.						
18. Slopes the test tube.						
19. Applies heat to the top of the water.						
20. Adjusts the clamp to the proper height.						
21. Clamps firmly but without excessive pressure.						
22. Rotates bottle when pouring.						
23. Estimates one teaspoonful.						
24. Mixes it on a piece of paper.						
25. Uses V paper to insert it in test tube.						
26. Twists stopper when inserting in test tube.						
27. Sets test tube horizontal.						
28. Twists glass into rubber tube.						
29. Tests water pressure before filling jar.						
30. Fills pan to suitable depth.						
31. Points overflow into sink.						
32. Uses glass to cover bottle when inverting.						
33. Allows no air to enter.						
34. Sets bottle on table inverted.						

APPENDIX F

PRACTICAL LABORATORY TEST

(Answer Sheet)

Test I . . . 1 . . . . . ( )

Test II . . . 2 . . . . . ( )

3 . . . ( )

Test III . . 4 . . . . . ( )

5 . . . . (c.c.)

6 . . . . (N.)

Test IV . . . 7 . . . . . ( )

8 . . . . . ( )

9 . . . . . ( )

Test V . . . 10 . . . . . ( )

11 . . . . . ( )

12 . . . . . ( )

Test VI . . 13

14 . . . . (gm.)

Test VII . . . . . 15 . . . ( )

Test VIII . . . . 16 . . . ( )

17 . . . ( )

18 . . . ( )

Test IX . . . . . 19 . . . ( )

20 . . . ( )

Test X . . . . . 21 . . . ( )

22 . . . ( )

Test XI 23 . . . ( )

24 . . . ( )

25 . . . ( )

Test XII . . . . . 26 . . . ( )

27 . . . ( )

Test XIII . . . . 28 (. . . test

29 for . . . .)

30 . . . . ( )

APPENDIX G

DATA

NAME	IQ	Criterion Test	Pencil & Paper Test	Notebooks	Teacher's Estimates	Criterion A	Criterion B
Duncan M.	120	50	37	140	75	24	26
Ratushny F.	115	47	40	148	87	24	23
Glaum L.	130	46	36	46	37	24	22
Greenough R.	101	45	35	90	72	24	21
Mah G.	118	45	39	110	63	22	23
Westlund W.	133	45	38	145	85	21	24
Costanzo P.	141	44	30	109	44	24	20
Gronlie M.	112	44	28	130	65	19	25
Scrimgeour G.	133	44	41	103	86	23	21
Gillingham J.	133	43	29	136	58	23	20
Lortie G.	101	42	32	109	18	23	19
Davies J.	121	40	35	82	58	19	21
Johanssen J.	111	40	27	137	70	19	21
Lum W.	108	40	29	121	66	19	21
Rosen L.	119	40	33	141	71	23	17
Wilson T.	147	40	39	103	63	17	23
Crane R.	126	39	27	127	66	19	20
Brown R.	127	38	37	127	76	17	21
Con B.	103	38	24	125	47	22	16
Hall J.	121	38	35	140	75	20	18
Lamb K.	149	38	36	135	82	21	17
Mitchell W.	131	37	37	136	81	19	18
Roscoe M.	114	36	34	123	61	14	22
Baker C.	100	35	29	113	33	19	16
Mitchell R.	123	35	31	133	78	17	18
Vea A.	129	35	26	102	60	16	19
Wong C.	107	35	30	136	66	22	13
Yip Y.	114	35	22	101	34	22	13
Campbell R.	110	34	35	104	51	17	17
Jarvis A.	126	34	24	102	50	18	16
Kraft D.	114	34	22	116	38	20	14
Moore R.	109	34	32	141	70	18	16
Ottewell D.	106	34	23	109	32	22	12
Tillyer D.	113	34	22	100	39	17	17
Brown D.	132	33	22	99	35	17	16
Dennis G.	94	33	28	117	65	17	16
Fortin L.	114	33	32	93	52	18	15
Carle R.	130	33	30	128	62	13	20
Lee N.	122	33	34	121	68	15	18
Baker G.	90	32	31	98	58	21	11
Bell H.	122	32	29	121	54	18	14
Carfrae M.	102	32	38	132	74	18	14
Chin R.	120	32	22	122	46	21	11
Knight R.	129	32	20	129	57	17	15

APPENDIX G (Cont'd.)

Name	IQ	Criterion Test	Pencil & Paper Test	Notebooks	Teacher's Estimates	Criterion A	Criterion B
Makort A.	113	32	30	46	59	19	13
Williams F.	117	32	22	105	24	20	12
Goff G.	106	31	37	118	31	14	17
Lee C.	114	31	28	111	63	17	14
Kihara S.	103	31	21	126	50	12	19
Bouzevetsky N.	116	30	28	77	33	14	16
Godson K.	127	30	19	109	55	20	10
Hendry P.	127	30	30	104	66	16	14
Welbourn C.	113	30	22	113	38	14	16
Yee B.	122	30	26	113	40	14	16
Borsato F.	90	29	32	144	80	15	14
Kisielewich P.	117	29	23	78	33	14	15
Lowe D.	120	29	19	123	67	18	11
Newton S.	115	29	25	112	30	13	16
Saimoto J.	88	29	21	125	37	13	16
Shynkaryk W.	109	29	25	100	50	13	16
Smith C.	90	29	15	111	10	20	9
Brisseau G.	119	27	22	125	59	16	11
HHenderson P.	117	27	23	128	54	14	13
Potter R.	123	27	23	55	6	16	11
Englemann M.	108	25	19	108	44	14	11
Lawrence W.	105	25	17	113	25	17	8
Oberholtzer B.	111	25	22	102	52	13	12
Shillington S.	115	25	27	122	26	14	11
Sweet D.	107	25	24	126	51	11	14
Perdia N.	91	23	19	105	67	12	11
Lessman E.	107	20	18	118	21	13	7
Smith J.	128	20	27	121	41	8	12
POSSIBLE SCORE	--	64	50	162	100	34	30

CRITERION TEST is composed of two parts; A. The Revised Horton Test - a test of manipulating apparatus, B. The Practical Test on the Laboratory Experiments of Chemistry 91.

PENCIL AND PAPER TEST is a written test of fifty items based on the criterion.

THE NOTEBOOK is the score on the first fifteen experiments in the student's notebook prior to the investigation.

THE TEACHER'S ESTIMATE is an estimated score of the student's ability to do laboratory work by his teacher, viz., the investigator.

APPENDIX H

INTERNAL CONSISTENCIES? VALIDITIES AND DIFFICULTIES OF ITEMS ON PENCIL AND PAPER TEST

Item	Internal Consistency	Validity Coefficient	Difficulty Index	Item	Internal Consistency	Validity Coefficient	Difficulty Index
1.	.00	-.25	.85	26.	.31	-.05	.49
2.	.31	.26	.50	27.	.24	.15	.76
3.	.68	.38	.83	28.	.00	.33	.75
4.	.15	-.30	.86	29.	.26	.05	.41
5.	.48	.40	.70	30.	.33	.26	.26
6.	.48	.16	.68	31.	.35	.33	.26
7.	.00	-.05	.53	32.	.36	.36	.52
8.	.15	.31	.78	33.	.21	-.07	.29
9.	.00	-.06	.68	34	.38	.23	.63
10.	.26	.44	.61	35.	.36	.21	.50
11.	.10	.41	.44	36	.10	.21	.44
12.	.35	.40	.75	37.	.31	.10	.43
13.	.60	.45	.60	38.	.00	-.40	.08
14.	.59	.51	.61	39.	.41	.22	.63
15.	.38	.51	.65	40.	.81	.65	.71
16.	.51	.47	.79	41.	.38	.33	.63
17.	.45	.38	.65	42.	.70	.55	.51
18.	.25	.13	.79	43.	.75	.59	.64
19.	.35	.21	.79	44.	.59	.58	.76
20.	.20	.24	.82	45.	.45	.28	.58
21.	.45	.21	.53	46.	.45	.28	.46
22.	.24	.00	.28	47.	.59	.20	.35
23.	.44	.65	.75	48.	.44	.44	.26
24.	.42	.38	.60	49.	.68	.60	.26
25.	.22	.33	.33	50.	.21	-.06	.26

These validities were determined from a table of values of the Product-moment Correlation in a normal Bivariate Population corresponding to given proportions of success, given by Thorndike[1] and prepared by the Cooperative Test Service from a chart by Flanagan. The upper and lower groups were determined on the basis of the scores on the Pencil and Paper Test.

1 Thorndike, R.L., Personnel Selection, New York: John Wiley and Sons, Inc., 1949, pp. 347-351.

APPENDIX I

INTERNAL CONSISTENCIES AND DIFFICULTIES OF CRITERION TEST ITEMS

Item	Coefficient	Difficulty	Item	Coefficient	Difficulty
1.	.27	.96	1.	.37	.46
2.	.55	.78	2.	-.23	.54
3.	.35	.72	3.	.21	.54
4.	.18	.87	4.	.63	.69
5.	.56	.54	5.	.60	.43
6.	.48	.74	6.	.68	.17
7.	.43	.95	7.	.55	.49
8.	.10	.89	8.	.50	.43
9.	.15	.54	9.	.21	.67
10.	.18	.75	10.	.00	.49
11.	.15	.40	11.	.00	.52
12.	-.10	.89	12.	.39	.68
13.	.25	.89	13.	.63	.54
14.	-.11	.36	14.	.52	.65
15.	.24	.73	15.	.12	.56
16.	.34	.33	16.	-.05	.43
17.	.33	.68	17.	.11	.29
18.	-.05	.89	18.	.30	.48
19.	.30	.74	19.	.28	.68
20.	.00	.20	20.	.15	.54
21.	.06	.33	21.	.48	.72
22.	.25	.06	22.	.71	.79
23.	.21	.38	23.	.50	.49
24.	.40	.22	24.	.07	.18
25.	.55	.78	25.	.16	.36
26.	.07	.24	26.	.42	.79
27.	.54	.17	27.	.67	.76
28.	.40	.35	28.	.11	.63
29.	-.15	.04	29.	.40	.70
30.	.40	.20	30.	.51	.76
31.	.48	.22			
32.	.59	.60			
33.	.36	.28			
34.	.68	.79			

These validities were determined from a table of values of the Product-moment Correlation in a normal Bivariate Population corresponding to given proportions of success, given by Thorndike[1] and prepared by the Cooperative Test Service from a chart by Flanagan. The upper and lower groups were determined on the basis of the scores on the criterion test.

1 op.cit., pp.347-351.

APPENDIX J

T-SCORES FOR THE PENCIL AND PAPER TEST

Raw Score	Percentile	T Score	Raw Score	Percentile	D Score
50		85	25	42	45
49		83	24	36	44
48		81	23	30	42
47		80	22	24	41
46		78	21	17	39
45	100	77	20	12	37
44	100	75	19	9	36
43	100	74	18	6	34
42	100	72	17	4	33
41	99	70	16	3	31
40	98	69	15	2	29
39	96	67	14	1	28
38	93	66	13	0.5	26
37	90	64	12	0	25
36	86	61	11		23
35	82	59	10		22
34	79	58	9		20
33	75	56	8		19
32	73	55	7		17
31	71	54	6		15
30	67	53	5		14
29	63	51	4		12
28	59	50	3		10
27	54	48	2		9
26	48	47	1		7
			0		6

Percentiles computed from graph prepared from frequency distribution.

The Derived scores were computed from the formula:

$$T.S. = \frac{10\,(X - M)}{S.D.} + 50$$

Where T.S. is the derived score.

X is the raw score.

M is the mean of the distribution, viz., 28.08.

S.D. is the standard deviation, viz., 6.316.

APPENDIX K

PENCIL AND PAPER TEST

SCALED TO PLACE FIFTEEN PERCENT BELOW A CRITICAL SCORE OF 50

Raw Score	Scaled Score	Raw Score	Scaled Score
10	10	30	66
11	11	31	68
12	12	32	70
13	20	33	72
14	25	34	74
15	30	35	76
16	34	36	77
17	38	37	79
18	41	38	82
19	45	39	86
20	46	40	90
21	48	41	92
22	50	42	95
23	52	43	97
24	54	44	97
25	56	45	98
26	59	46	98
27	61	47	99
28	63	48	99
29	65	49	99
		50	100

The Scaled Score was derived from cumulative frequency curves based on (1) the raw scores of the pencil and paper test, and (2) a normal distribution of scores with the median set at 63 and the standard deviation set at 13. This method is employed by the British Columbia Department of Education in scaling scores on University Entrance Examinations.