

DIFFERENTIAL RETENTION OF
COURSE OUTCOMES IN EDUCATIONAL PSYCHOLOGY

by
William P. McDougall

A DISSERTATION

Presented to the Faculty of
The University of Nebraska in the Teachers College
in Partial Fulfillment of Requirements
For the Degree of Doctor of Education
Department of Educational Psychology and Measurements

Under the Supervision of Professor Charles O. Heidt

Lincoln, Nebraska

1957

TITLE

DIFFERENTIAL ESTIMATION OF COURSE OUTCOMES

IN EDUCATIONAL PSYCHOLOGY

BY

William Phillips McDougall

APPROVED

DATE

Charles O. Neidt

July 26, 1957

Joseph L. French

July 26, 1957

Madison Brewer

July 26, 1957

Howard E. Tempero

July 26, 1957

M. F. Thorpe

July 26, 1957

SUPERVISORY COMMITTEE

ACKNOWLEDGMENTS

The author is very grateful to his major advisor, Professor Charles O. Heidt, Chairman, Department of Educational Psychology and Measurements, for his generous assistance and inspiration in the development and preparation of this thesis. Appreciation is also expressed to Dr. Robert T. Littrell who made the experiment possible by taking a major part of the responsibility for the collection and processing of the test data.

The author also wishes to thank the instructors in the Department of Educational Psychology and Measurements, Elementary and Secondary Education, for their cooperation by allowing class time for the collection of the data.

W.P.H.

TABLE OF CONTENTS

| | Page |
|---|------|
| ACKNOWLEDGMENTS | i |
| TABLE OF CONTENTS | ii |
| LIST OF TABLES | iv |
| LIST OF FIGURES | v |
| | |
| I INTRODUCTION | 1 |
| A. Statement of Problem | 1 |
| B. Review of Related Literature | 3 |
| 1. Instructional Objectives | 3 |
| 2. Retention Studies | 10 |
| C. Need of Study | 15 |
| | |
| II PLAN AND PROCEDURE | 17 |
| A. Trial Test Construction | 17 |
| 1. How the Trial Test was Constructed | 18 |
| 2. Description of the Test Items | 19 |
| a. Knowledge | 20 |
| b. Comprehension | 21 |
| Translation | 21 |
| Interpretation | 22 |
| Extrapolation | 23 |
| B. The Trial Test | 24 |
| 1. The Trial Test Group | 24 |
| 2. Trial Test Administration | 24 |
| C. Refinement of the Test | 25 |
| 1. Item Analysis | 25 |
| 2. Evidence of Reliability | 26 |
| 3. Homogeneity of Test Behavior | 27 |
| D. Administration of the Refined Test | 29 |
| 1. Description of the Test | 29 |
| 2. Description of the Subjects | 30 |
| E. Treatment of the Data | 31 |
| 1. Item Difficulty and Discrimination | 31 |
| 2. Reliability of the Tests | 31 |
| 3. Homogeneity of Tests | 32 |
| 4. Evidence of Validity | 32 |
| 5. Retention | 33 |
| | |
| III FINDINGS OF THE STUDY | 37 |
| A. Analysis of the Tests | 37 |
| 1. Item Difficulty | 37 |
| 2. Item Discrimination | 38 |
| 3. Evidence of Reliability | 41 |
| 4. Homogeneity of Tests | 42 |
| 5. Evidence of Validity | 46 |

| | | |
|----|--|-----|
| B. | The Study of Retention | 47 |
| 1. | The Retention Group | 48 |
| 2. | Relationships Between Pretest - Test - Retest | 49 |
| 3. | Differences in Average Performance for the Various Test Administrations | 50 |
| 4. | The Retention Graph | 51 |
| 5. | Average Percent of Gain Retained | 52 |
| 6. | Occurrence of Certain Patterns on the Successive Administrations of the Test Items. | 54 |
| IV | IMPLICATIONS OF THE INVESTIGATION | 56 |
| V | SUMMARY | 59 |
| | SELECTED REFERENCES | 63 |
| | APPENDIX A | 66 |
| | APPENDIX B | 85 |
| | APPENDIX C | 88 |
| | APPENDIX D | 100 |
| | APPENDIX E | 101 |

LIST OF TABLES

| | | Page |
|----------|--|------|
| Table 1 | Means, Standard Deviations and Spearman-Brown Estimates of Reliability for Tests | 27 |
| Table 2 | Intra- and Inter-Correlations Between Tests | 28 |
| Table 3 | Values of F Between Tests for Departure from Homogeneity | 29 |
| Table 4 | Item Difficulty and Item Discrimination for Test Score and Combined Total Score | 39 |
| Table 5 | Spearman-Brown and Kuder-Richardson Estimates of Reliability | 42 |
| Table 6 | Intra- and Inter-Area Correlations Between Tests | 44 |
| Table 7 | Values of F for Tests of Homogeneity Between Tests | 44 |
| Table 8 | Values of F for Tests of Homogeneity Between Tests (Interpretation-Extrapolation Combined) | 45 |
| Table 9 | Correlation Coefficients Between the Three Administrations of the Tests | 50 |
| Table 10 | Differences in Mean Scores and the Accompanying t-value Between the Three Administrations of the Tests | 51 |

LIST OF FIGURES

| | Page |
|---|------|
| Figure I Retention Graph of the Three Tests: Knowledge, Translation, Interpretation-Extrapolation | 52 |

CHAPTER I

INTRODUCTIONA. Statement of Problem

Tyler stated: "...the organization of courses and the development of examinations should center about those objectives having more permanent values."¹(35:203)¹ Of major concern to the classroom teacher is the need for accurate evidence concerning the degree of permanency of each instructional objective. Such information would certainly enable the teacher to function more effectively in guiding the learning activities of students.

The relative permanency of different types of learning has often been overlooked. Frequently it has been assumed that the results of the usual instructional procedures are of lasting value and that the attainment of one objective can be inferred from measured attainment of another objective. The practices of defining each educational goal, validating evaluation devices against defined aims, and studying the attainment of each one of these aims to determine its degree of permanency are clearly needs in our instructional programs.

The purpose of this study was to measure retention of different course outcomes in a beginning course in educational psychology. The outcomes examined included knowledge and the intellectual skills and abilities specified as translation, interpretation and extrapolation. These objectives were defined by the "Taxonomy of Educational Objectives"⁽⁶⁾,

1. The number which appears first in the parenthesis refers to a title in the list of references. The number following the colon indicates the pages on which the citation appears.

and are explained briefly in the following paragraph. The subject matter content studied was delimited to one area, tests and measurements, to permit more intensive and uniform sampling of the different course outcomes. The samples of responses to the test situations were taken from the Education 62 classes at the University of Nebraska. (Education 62, Human Behavior and Development, is the second of two courses in educational psychology taken by teacher trainees.) Follow up testing was done approximately a semester later in the teacher training sequence.

According to the Taxonomy of Objectives, six major hierarchical levels of educational objectives are defined in the cognitive domain. Of these six the first two, specified as knowledge and comprehension probably include the largest general classes of intellectual activities emphasized in schools and colleges. (6:89) For this reason these two major categories were selected for this study. The category of knowledge includes recognition and recall of specifics, ways and means of dealing with specifics, and knowledge of universals and abstractions. Test items designed to measure these categories seem to be very similar and, therefore, all knowledge items were defined as being in a single category. The comprehension level includes items designed to measure the ability to translate knowledge from one form to another, to interpret data and to extrapolate from data. Items designed to measure the sub-categories on the comprehension level were grouped as separate tests. A more complete description of these behavioral objectives as defined in the Taxonomy is given in Chapter II in connection with the trial test construction.

The first task undertaken in this study was the construction and validation of the tests to be used to measure the desired outcomes. These tests were constructed and analyzed both in trial and final test form

using the following criteria:

1. The relationship of each test item to explicitly defined objectives.
2. The contribution of individual items to the effectiveness of its respective test.
3. The degree to which the separate tests measure the desired behavioral functions.
4. The degree to which the instrument satisfies conditions of empirical validity and reliability.

The second major part of the study consisted of the use of the tests to determine the degree to which the separately evaluated abilities were retained. For this purpose the tests were administered at three times: as a pretest at the beginning of the course, a post test as part of the final examination and again as a retention test approximately four months later. The differences between the sets of scores were then examined to ascertain the degree of permanency of each of these course outcomes.

B. Review of Related Research

1. Instructional Objectives

Examination of the literature clearly points out the need for careful planning and formulation of objectives in the construction of achievement tests. A summary of steps suggested by Tyler emphasizes this point. The steps are as follows: (24:5)

- 1) Formulation of course objectives
- 2) Definition of each objective in terms of student behavior
- 3) Collection of situations in which students will reveal presence or absence of each behavior

- 4) Presentation of situations to the students
- 5) Evaluation of students' reaction in light of each objective
- 6) Determination of objectivity of evaluation
- 7) Improvement of objectivity when necessary
- 8) Determination of reliability
- 9) Improvement of reliability when necessary
- 10) Development of more practical methods of measurement when necessary

The importance of objectives has been effectively stressed in the writing of Peters (26:148) who showed that few tests were validated against educational objectives and suggested that they should be derived on this basis. Jordan also emphasized this point of view (20:5), and stressed the necessity that the objectives of education be clearly defined or else the appropriate measuring instruments cannot be constructed. Other writers such as Bean (4:17), and Vaughn (22:160) have stressed the basic importance of delineated objectives.

Vaughn stated that too often a test outline is confined to subject matter alone and the type of behavior to be examined is left to the judgment of an inexperienced item writer. He indicated that a test outline can be used to show clearly not only what different areas of subject matter are to be covered but also the types of behavior to be elicited with respect to each area. Rorters (28:27) concurred with this point of emphasis and stated that the distinction between general and specific objectives should be made to facilitate the comprehension of objectives. He further pointed out that objectives should be stated in terms of observable changes in pupil behavior. Likewise, Tyler (53:30) emphasized that the most useful form for stating objectives is to express them in

terms which identify both the kind of behavior to be developed in the student and the content area of life in which this behavior is to operate. He stated that clearly formulated objectives included both the behavioral aspect and the content aspect. Tyler pointed out that objectives are often stated in terms of things which the instructor is to do, lists of topics, concepts, generalizations, or other elements of content to be dealt with, or as generalized patterns of behavior which failed to indicate more specifically the area of life or the content to which the behavior applied. He further stated that since the real purpose of education is to bring about changes in the students' pattern of behavior, it becomes clear that the statement of objectives should be in terms of behavioral change.

Flanagan (11) proposed that a "method of rationales" be employed for the purpose of clear and precise definitions of what is to be measured. The method begins with a list of behaviors to be sampled or predicted. The development of these rationales consists of three parts.

- 1) Description of behavior
- 2) Analysis of behavior
- 3) Formulation of item specifications

Description of behavior involves the definition, delimitation and illustration of the variety and scope of the actions included. Analysis of behavior includes clarifying it with respect to other behaviors and making inferences about its nature, culminating in the formulation of one or more hypotheses regarding its generality and predictability. Formulation of item specification carries the procedure on to describing a specific type of item which, it appears, should provide a valid estimation of the specified behavior.

Micheels (25:92) also emphasized the importance of the behavioral aspect of objectives. He proposed four steps to follow in setting forth the test objectives. The steps are:

- 1) List the major objectives for which appraisal is desired.
- 2) Examine the course content for additional objectives.
- 3) Analyze and define each objective in terms of expected student outcomes.
- 4) Establish a table of specifications for the tests.

He suggested that step three is the inventory step. Various elements are listed that are a part of each objective and meaning is given to each element by defining it in terms of student behavior.

Much attention has been given in the literature to the listing and categorization of the broad general outcomes of the educative process.

(2:14) Most frequent among the outcomes listed are such types as skills, knowledges, concepts, understandings, applications, attitudes, interests and adjustments.(16:168) (14:16)

In contrast to plans involving broad general classifications, other plans have been proposed which require more specific definition of objectives. The report, "A Design for General Education"(2:31), illustrated how the objectives in various fields can be classified into three basic categories: 1) knowledge and understanding, 2) skills and abilities, and 3) attitudes and appreciations. The authors of this report asserted that the three basic categories could be used as subdivisions for more general educational outcomes. The objectives for general education are organized under ten broad outcomes and in turn these are defined more specifically under the above listed categories. Approximately two hundred sub-objectives presented are useful in defining more specifically subject matter content and behaviors involved in general education.

Tyler illustrated the use of a two dimensional chart in stating objectives for a course (33:32) in biological science. One dimension of the chart defines the content aspect of the objectives, the other the behavioral. There are seven types of behaviors specified in the biological science course used as an illustration. Tyler describes these as follows:

"The first type of behavior is to develop understanding of important facts and principles. The second type is to develop familiarity with dependable sources of information... The third type of behavior is to develop ability to interpret data - that is, to draw reasonable generalizations from the kinds of scientific data likely to arise in this field. The fourth type of behavior is to develop ability to apply principles that are taught in biological science to concrete biological problems that arise in everyday life - hence to be able to carry on problem-solving activities in this field. The fifth type of behavior is to develop the ability to study and report the results of study. The sixth is to develop broad and mature interests as they relate to biological science, and the seventh is to develop social rather than selfish attitudes in this area."

The chart also includes a statement of the content aspects of the activities. Examples of these are such headings as nutrition of human organisms, digestion, circulation, respiration and reproduction. The course is then viewed as developing various sorts of behavior in relation to these aspects of content.

The "Taxonomy of Educational Objectives" includes the most careful definitions of objectives yet made. (6) Over several years a group of college and university examiners attacked the problem of creating a comprehensive classification of educational objectives. (6:4) Several uses are proposed for such a classification. The device is intended to facilitate communication among teachers and examiners, provide a framework upon which to plan learning experiences and prepare evaluation devices as well as to serve other useful educational needs. The efforts of this project resulted in the publication of a Handbook #1, which consists of a detailed

logical and psychological classification of educational aims in the cognitive domain. As the classification is now organized it contains six major classes: (6:10)

- 1.00 Knowledge
- 2.00 Comprehension
- 3.00 Application
- 4.00 Analysis
- 5.00 Synthesis
- 6.00 Evaluation

The authors conceive of these classes as representing a general hierarchical order. The behaviors of one class are likely to make use of and be built on behaviors found in the preceding classes. Since extensive use was made of the Taxonomy in the present study it will be described in greater detail in Chapter II.

Whereas the foregoing literature pertains to the general process of establishing objectives, authors of other publications stress the importance of clearly specifying the type of objectives being measured. Tyler (21:6) has studied the relationships between the recall of information on tests, the application of principles and the ability to draw inferences. These three types of tests were administered to sixteen different college classes in eleven different subject matter areas. The average number of students in the classes was 217 and ranged from twenty-two to 624. The resulting correlations between recall and application items ranged from .31 to .50 and were centered about .45. Correlations between the recall and inference items ranged from .27 to .60, averaging about .35. The ability to apply principles and to draw inferences yielded an average correlation of .40, ranging from .33 to .54. Tyler concluded that students did not

develop corresponding degrees of achievement in mere recall and in achievement in such higher mental processes such as the application of principles and the ability to draw inferences.

The results of McConnell's study (24:78) at the University of Minnesota agree with the work done by Tyler. McConnell dealt with tests in three subject matter divisions and achievement of three kinds of objectives; namely, knowledge of vocabulary, knowledge of facts and principles and ability to apply facts and principles. His analysis by the method of intercorrelation yielded an average correlation of .66. He also studied the extent of differential measurement by the method of item discrimination.

The assumption underlying this procedure was that an item designed to discriminate between students of high and low ability on one test would do so best when its own test total was used as the criterion rather than the total of a test designed to measure another objective. He found, using this method, that the average of all correlations within the three subject matter areas was .87 and the average among the three divisions was .66. This difference is statistically significant and means that there is relatively greater homogeneity of behavior within objectives than among objective groupings.

Other similar studies have resulted in comparable findings. Dodell (5) using a sample of 324 pupils in ninth grade general science found a correlation of .65 between performance on items requiring students to recall and to infer from science principles. Johnson (19) reported slightly higher coefficients, averaging about .85, between knowledge of fact and principles and the ability to apply them to new situations in college science courses. Brown (25:50), in the area of home economics, reported

that the correlations between a knowledge of scientific principles of cookery and the quality of food cooked or ability of people to manage their work was less than .50. Horrocks (16) studied the relationship of knowledge of facts and principles and the ability to apply them in the area of adolescent development. He utilized a criterion test of knowledge and three case study tests of application in testing three hundred upper classmen and graduate students. Each of the case study tests dealt with a different problem. The problems were in the social, academic and emotional areas. Resulting correlations between the criterion test and each of the case studies were .46, .41 and .26, respectively.

From the foregoing studies it is apparent that achievement of one objective cannot be inferred from the achievement of another. Remmers has expressed this point when he concluded: (28:31)

"...the educator must clearly define each objective in terms of the measure of its attainment. The attainment of a particular objective cannot be inferred from measured attainment of another objective."

2. Retention Studies

Though a large number of retention studies for many different school subjects have been reported, few have dealt directly with the problem of measuring the relative permanency of different kinds of course outcomes. In general, the studies show that forgetting proceeds rapidly at first and then more slowly. Much of what is shown by the usual final examination at the end of a college course is forgotten within four to six months. (27:544) Permanency of some outcomes seems to be, however, much greater than others. The first few studies will indicate the general nature of retention in college courses. The latter studies will treat the problem of differential permanency of instructional objectives.

A study of the retention of information learned in college courses was reported by Greene. (15:262) Retention of information demonstrated on the final examinations was studied in zoology, psychology, and physiological chemistry at intervals ranging from four to twenty months. The content of the final examination placed emphasis upon the recognition and recall of specific information rather than problem solving or logical organization of material. The final examination, which was given in June, was readministered to thirteen students in zoology, twenty-six in psychology and eighty-eight in chemistry during the following October. Allowing for initial learning, approximately one-half of the material reported correctly on the June final was lost in the four-month period. The proportion of the final examination scores which was retained eight months later averaged roughly one-quarter for zoology and one-fifth for psychology, and at the end of twenty months this averaged from one-tenth to one-fifth. Chemistry was not studied beyond the four-month period. Greene also reported correlations between the June and October scores for fifteen subjects in zoology, twenty-six in psychology and eighty-eight in chemistry. The resulting correlations were .708, .412 and .434, respectively. He suggests that the higher correlation in zoology may be attributed to the fact that it was a laboratory course and may have resulted in more consistent retention than a lecture course.

Watson (40) studied the permanence of material learned in a required introductory psychology course. One hundred college students were tested by typical examination questions for immediate recognition and recall and then, having been divided into six highly comparable subgroups ranging in size from ten to thirty-three, were tested for delayed retention after intervals of two, four, six, eight, ten, eighteen, twenty, twenty-two,

thirty, thirty-two, thirty-four, forty-two, forty-four, forty-six, fifty-four, fifty-six and fifty-eight months. Each subgroup was tested for three delay periods, a test of different type being used for each interval. Watson concluded that although forgetting increased with time, the point of complete forgetting was not reached even after fifty-eight months. Different results were obtained for recognition and recall type items. After a delay of two and one-half months he reported that ninety-seven percent retention occurred on the recognition scores; after a period of twelve months eighty-four percent occurred. The material learned that required recall in the test situations yielded much lower percentages. For the two and one-half month delay the percentage was thirty-seven, for six months thirty-four, and for nine months thirty-three percent. In general, the recognition curves decreased gradually and progressively and the recall curves decreased abruptly and progressively throughout the delay period.

A study of retention of knowledge acquired in a course in general psychology reported by Lurich (9) showed a higher degree of permanence of learning. Two final examinations constructed by the Department of Psychology at the University of Minnesota were used separately to study retention over periods of six and nine months. The one examination consisting of single choice, analogy, wrong word and completion type items was given at the completion of a psychology course and again nine months later to ninety-nine students. The other similar test contained no completion items but did include matching items and was given as a retest to eighty-three individuals six months later. The tests apparently measured largely factual information but this cannot be determined from the report. The mean score nine months after course completion was approximately

seventy-five percent of the mean at the close of the course. On the other retest after the six-month period, the mean score was ninety percent of the mean at the close of the course. Murich concluded that retests in psychology show that students retain a substantial amount of the measured knowledge six and nine months after they have taken the course.

The following studies will review findings where attempts have been made to identify separately the permanency of different types of course outcomes.

Tyler (35) studied the scores of eighty-two typical students in a course in elementary zoology who had been given several types of examination exercises at the close of the course and again fifteen months later. Five types of performance were measured. These were: naming animal structures pictured; identifying technical terms; recalling factual information; applying principles to new situations; and interpreting new experiments. Seventy-seven percent of the gains made during the course in the ability to name animal structures was lost in the fifteen-month period following. For the abilities to identify technical terms and recall information, approximately one-fourth of the gain made was lost. The ability to apply zoological principles to new situations or to interpret data from experiments showed no loss over the fifteen-month retention period even though these students had taken no other course in zoology during this time.

An experiment by Hart (41), also in zoology, yielded similar results. Separately measured course objectives were studied over a period of three years. The ability to apply principles and to interpret new experiments demonstrated in three years a percentage increase, fifty-eight and nineteen percent, respectively, over the students' performance at the end of

the course. There was a loss in the gain made during the course of over fifty percent in ability to remember terminology, functions of structures and main ideas. Over eighty percent loss occurred in associating names with structures. Wert's study showed that retention was greatest in areas involving application and interpretation and least in areas involving informational objectives.

Five different objectives in high school chemistry were chosen by Frutchey (12) to be studied. These objectives were: selection of facts; application of principles; terminology; symbols, formulas, valence and balancing equations. Tests measuring these objectives were given to an average of eighty-three students as a pretest, again nine months later at the end of the course and a third time one year after completion of the course. The retention results were reported in terms of the percent of gain made in the course that was retained. For selection of facts eighty-four percent was retained, ninety-two percent for application of principles, sixty-six percent for terminology and about seventy-one percent for symbols, formulas, valence and balancing equations. Frutchey concluded that retention was greatest in the more general types of behavior.

Differential retention was also reported by Tyler in ninth grade general science.(36) Sixty-eight pupils were given the Rich-Poponee test, largely a measure of information, a multiple-choice explanation test, and a test requiring students to generalize from given facts. The tests were repeated eight months after the completion of the course. The mean loss for the Rich-Poponee was 11.1, for the explanations test 1.8, and for the test requiring generalizations 1.2. Apparently the ability to explain and generalize from known facts was retained better than mere information.

Individual differences in retention of general science subject matter involving recall of factual information, ability to explain scientific phenomena and ability to draw conclusions from given data were reported by Ward. (39) The tests measuring each of these behavioral outcomes were administered in June and again in September, approximately three and one-half months later to sixty-three students. The percentage of forgetting during this period was greatest for the factual information and conclusion parts, being approximately seventeen percent relative loss. For the test requiring the students to explain scientific phenomena, the retention was greatest with a mean loss of 9.1 percent. Ward felt that the findings of this study strongly suggest that ability to apply principles and to explain phenomena, and problem solving procedures, are retained over a long period with slight loss. He concluded also that the results reinforce a commonly accepted belief among educators that the permanent outcomes of teaching are to be found among the so-called "intangible objectives."

C. Need of Study

The need of the present study is suggested by the following quotation from the Taxonomy. (6:23)

"For the most part, research on problems in retention, growth and transfer has not been very specific with respect to the particular behavior involved. Thus, we are usually not able to determine from this research whether one kind of behavior is retained for a longer period of time than another or which kinds of educative experiences are most efficient in producing a particular kind of behavior. Many claims have been made for different educational procedures, particularly in relation to permanence of learning; but seldom have these been buttressed by research findings."

Judd has said, (21:4) "If by any means the educational system can discover how to promote even in the slightest measure the development of

the higher mental processes, great advantage will be gained for civilization." A first step in giving prominence to the so-called higher processes is to identify clearly which of these types of learning are of most lasting value. The review of literature in the previous section certainly strongly suggests such abilities as making inferences and applications stand out as demonstrating greater degrees of permanence than more factual information. The need to study other instructional outcomes such as those being studied in this investigation then becomes increasingly evident. In addition, the Taxonomy becomes a highly useful instrument to standardize the description of research findings and to facilitate communication of these results. Such behaviors as are studied can be much more clearly and universally identified, a task that has heretofore been difficult.

CHAPTER II

PLAN AND PROCEDURE

The general plan of this study involved the construction of tests to measure a variety of educational objectives in the content area of test and measurements. These tests were then used to study the relative retention of different abilities. More specifically, items were designed to measure each of the following:

- 1) the ability to recognize or recall basic knowledge
- 2) the ability to translate this knowledge from one form to another
- 3) the ability to interpret data
- 4) the ability to extrapolate from data

These tests were then administered as a trial test to a group of educational psychology students who had completed units on tests and measurements. The tests were then refined and used as instruments to study the retention of the above mentioned abilities. The refined test was given as a pretest¹, a test at the completion of the course and a re-test approximately a semester later.

A. Trial Test Construction

The trial test contained approximately thirty multiple-choice items in each of four categories - knowledge, translation, interpretation and extrapolation.

1. For purposes of this study, "pretest" will be used to describe the administration of the test at the beginning of the unit, "test" the administration at the end of the unit and "retest" the administration at the end of the four-month retention period.

1. How the Trial Test was Constructed

The staff in the Department of Educational Psychology and Measurements has worked cooperatively in the development of a syllabus outlining the content to be included in Education 62, Human Behavior and Development. The syllabus has fifteen lessons, each containing important concepts and principles together with selected questions and readings to aid in their study.

Each of these lessons was analyzed, and the listed readings and basic materials in the area of tests and measurements were noted. Several lessons, particularly those on evaluation, contained a large number of measurement concepts and thus provided the major points of contact between the test items and the course. Attempts were not made to sample adequately all measurement concepts in Education 62 but rather the syllabus was used to insure the investigator that the content of each test situation was covered in the course.

It may be inferred from the introductory pages of the syllabus that emphasis in this course is not confined to knowledge alone but rather that such intellectual abilities and skills as understanding and application also receive considerable emphasis. The following quotation is taken from the syllabus:²

²The course is aimed at helping students to acquire a well-organized body of sound principles which will guide them in their efforts to understand behavior and wisely to influence behavior. The emphasis here is upon the expression, body of sound principles. A handful of untested beliefs about behavior is not psychology. Nor does psychology consist of a haphazard collection of facts about behavior. The aim of

2. The syllabus is a course guide developed by the staff in the Department of Educational Psychology and Measurements at the University of Nebraska and is used by all instructors in teaching Education 62, Human Behavior and Development. Copies may be obtained from this department.

Education 61-62 is to help students to develop a system of trustworthy ideas about the nature of human behavior and development and how best to put these to use in their work as teachers."

Though specific behavioral objectives are not enumerated in the above quotation, it should be safe to assume that a thorough comprehension of the subject matter is implied.

The Taxonomy was then employed and test items were constructed patterned after the definitions set forth in this classification. Items were devised to measure knowledge and were combined into a single test. On the level of behavior specified as comprehension, three different abilities and skills were considered, resulting in separate tests designed to measure translation, interpretation and extrapolation.

Further to insure the investigator that the content of each test item used paralleled the coverage in Education 62, each of four instructors teaching this course were asked to read and criticize the original list of test items. From the combined comments all materials and test inclusions thought to be inappropriate by any instructor were excluded from the original list of items. The resulting trial test contained 116 items and required approximately two hours for administration. Thirty of these items were designed to measure knowledge, thirty-two translation, twenty-six interpretation and thirty extrapolation. A more complete description of these objectives and test situations follows.

2. Description of the Test Items

In this section the Taxonomy definition will be given for each behavioral objective being tested, together with illustrative test items. A complete copy of the tests may be found in Appendix A.

a. Knowledge

"Knowledge, as defined here, involves the recall of specifics and universals, the recall of methods and processes, or the recall of a pattern, structure, or setting. For measurement purposes the recall situation involves little more than bringing to mind the appropriate material. Although some alteration of the material may be required, this is a relatively minor part of the task. The knowledge objectives emphasize most the psychological processes of remembering. The process of relating is also involved in that a knowledge test situation requires the organization and reorganization of a problem such that it will furnish the appropriate signals and cues for the information and knowledge the individual possesses. To use an analogy, if one thinks of the mind as a file, the problem in a knowledge test situation is that of finding in the problem or task the appropriate signals, cues and clues which will most effectively bring out whatever knowledge is filed or stored."(6:201)

The Taxonomy specifies several categories of knowledge including knowledge of specifics, knowledge of ways and means of dealing with specifics and knowledge of universals and abstractions in a field. Each of these categories is represented in the test. An example of a question requiring specific knowledge is as follows:

18. Standard achievement tests are most often reported in:
- (1) Standard scores
 - (2) raw scores
 - (3) grade placement scores
 - (4) quotient scores

An example of a question requiring knowledge of ways and means of dealing with specifics would be:

16. Which of the following best describes the accepted procedure in the use of I.Q. test results?
- (1) give the I.Q. to the parents and student if they request it and seem serious about the matter.
 - (2) never reveal the I.Q. to anybody
 - (3) reveal the I.Q. to the parents but not the student
 - (4) reveal the interpretation of the I.Q. to the parents or student

An example of a question requiring knowledge of universals and abstractions in a field is:

4. A generalization that might be made about most standard tests is that they:
- (1) are difficult for the teacher to administer
 - (2) are relatively inappropriate for most things we do in school
 - (3) are misleading if treated as the sole evidence of merit
 - (4) usually require more time than can be justified as a part of any single course

b. Comprehension

"This represents the lowest level of understanding. It refers to a type of understanding or apprehension such that the individual knows what is being communicated and can make use of the material or idea being communicated without necessarily relating it to other material or seeing its fullest implications."
(6:204)

The Taxonomy distinguishes three kinds of skills and abilities involved in comprehension:

Translation

"Comprehension as evidenced by the care and accuracy with which the communication is paraphrased or rendered from one language or form of communication to another. Translation is judged on the basis of faithfulness and accuracy, that is, on the extent to which the material in the original communication is preserved although the form of the communication has been altered."
(6:204)

Examples of questions constructed to measure translation are:

40. A major use of testing is for diagnosis. Which of the following test situations represents the best example of the foregoing statement?
- (1) a comprehensive achievement battery at the end of high school
 - (2) an achievement battery given early in the year
 - (3) an intelligence test
 - (4) a series of tests used to determine a student's grade
41. If Bill scored at the 88th percentile in Social Service on the Kuder Preference Test, it would indicate that:
- (1) Bill got 88% of the answers correct
 - (2) he has more ability in Social Service than 88% of his norm group
 - (3) only 12% of the norm group showed more interest in Social Service than he did
 - (4) that 88 out of 100 will do better than he did on this test

The first exercise involves translation of a formal statement by requiring the student to identify a concrete example. The second item involves the translation of quantitative data to its corresponding verbal meaning.

Interpretation

"The explanation or summarization of a communication. Whereas translation involves an objective part-for-part rendering of a communication, interpretation involved a re-ordering, rearrangement, or a new view of the material."
(6:205)

Examples of interpretation items would be:

Data are given below on five pupils enrolled in a class of thirty ninth graders. The test data are based on performance at the end of the first semester. Read over the summary and then show to which pupil each statement best fits by marking the pupil's number on the answer sheet.

| <u>Pupil</u> | <u>I.Q.</u> | <u>Calif. Ach. Test Performance</u> | | | <u>Teacher's estimate of Ach. Rank in Class</u> |
|--------------|-------------|-------------------------------------|--------------|--------------|---|
| | | <u>Arith.</u> | <u>Read.</u> | <u>Lang.</u> | |
| 1 | 88 | 9.1 | 8.0 | 8.3 | 20 |
| 2 | 99 | 9.7 | 9.6 | 9.5 | 14 |
| 3 | 132 | 9.5 | 9.8 | 10.2 | 12 |
| 4 | 138 | 11.8 | 12.3 | 12.0 | 3 |
| 5 | 101 | 10.0 | 10.1 | 10.9 | 4 |

42. The pupil who should be doing considerably better in his school achievement.
43. The accuracy of the I.Q. seems most doubtful in which case?
44. A bright student making good use of his ability.
45. Teacher regards abilities too highly according to test results.
46. Teacher's rank most consistent with test scores.

Each of the foregoing situations involves the ability to deal with a configuration of ideas or data recognizing the relation and relative importance of each. The inferences or generalizations made from the data do not extend beyond the data but are confined to the material presented.

Extrapolation

"The extension of trends or tendencies beyond the given data to determine implications, consequences, corollaries, effects, etc., which are in accordance with the conditions described in the original communication." (6:205)

Examples of extrapolation would be:

The five students for whom the data are given below are in kindergarten. These test data are based on test performance at the beginning of the second semester. After examining the data indicate which pupil best fits each of the following statements by marking the number of the student on the answer sheet.

| <u>Student</u> | <u>C.A.</u> | <u>I.A. on Stanford</u> <u>Binet</u> | <u>Percentile Rank on</u> <u>Readiness Test</u> |
|----------------|-------------|---|--|
| 1 | 5-10 | 7-4 | 72 |
| 2 | 6-4 | 5-4 | 22 |
| 3 | 5-10 | 5-5 | 64 |
| 4 | 5-8 | 5-6 | 45 |
| 5 | 5-6 | 6-10 | 38 |

Which student:

48. Is apparently in need of stimulating experiences but has fairly high aptitude?
50. Apparently comes from a very stimulating environment?
51. Is most characteristic of the average for this group?
52. Can you predict will have the lowest ability three years from this time?

The first two situations require the student to extend the implications of the data to another topic or situation. The third situation requires extension from a sample to a universe. The last item involves time dimension and requires prediction on the basis of the data presented.

To assure the investigator that each test item corresponded to the definitions as set forth in the Taxonomy, pooled judgments of several people were involved and whenever any doubt was evident about the nature of classification of the item it was revised or eliminated. The Taxonomy itself provides extensive definitions and descriptions of each of these

specified behaviors together with illustrative test items. Such a device enables test builders clearly to identify the process or behavior being measured.

B. The Trial Test

1. The Trial Test Group

Seventy-five students who were enrolled in Education 62 during the summer of 1956 comprised the trial test group. The group was fairly typical with the exception that a slightly higher percentage of students with teaching experience were enrolled at this time.

The classes in which the trial test was administered were using the same syllabus and text as were used later by the groups studied in the retention test. All units containing material in the test had been completed at the time of the test.

2. Trial Test Administration

The trial test was given by the respective instructors in Education 62 during two successive periods immediately prior to the final examination. The students were informed that part of the test would receive some weight in determining their final grade. They were not informed which items would be used for final grading and thus adequate testing motivation was assured.

The test was administered as a power test, i.e. ample time was provided for all members to attempt all items. In the case of a few members who could not finish within the two periods, additional time was allotted during a third period or by special arrangement.

C. Refinement of the Test

1. Item Analysis

The items in each test were examined for difficulty and discrimination ability. The test item difficulty, in terms of the percentage of the group who responded correctly to each item, is reported in Appendix B. The average level of difficulty for the knowledge pretest was approximately sixty-seven percent. The remaining three tests averaged about fifty-five percent difficulty. Items ranged in level of difficulty on all tests from approximately twenty percent to above ninety percent and clustering about the averages.

Item discrimination was determined by correlating each item with the total test score. To obtain these correlations the upper and lower twenty-seven percent of the distribution designated as the criterion variable were first identified. The proportions were then entered in an item analysis table (10) to identify the appropriate estimated correlation coefficients. Such coefficients indicate the tendency for students who make high scores on the total test to mark the individual item correctly. These correlations are also reported for each item in Appendix B.

To appraise items with regard to the item-criterion correlations it was first necessary to eliminate items in each test exceeding the standard error of the test correlations. The formula for the standard error is: (42:296)

$$S.E. = \frac{1}{\sqrt{N-3}}$$

where S.E. is the standard error of a correlation coefficient and N is the number of cases taking the test. Substituting in this formula,

$$S.E. = \frac{1}{\sqrt{15-3}}$$

the resulting standard error was .12. On this basis six items were eliminated from the knowledge test, two from translation, two from interpretation and four from the extrapolation test. These items were not used in any further analysis.

The appraisal of the remaining items was not done solely on the basis of these statistics but consideration was given to logical analysis of the item content. An attempt was made to select the items which measured a different aspect of content.

Additional attention was given to the foils or distractors of the items. An individual count was made to determine the frequency of use of each separate foil. From this count the weak distractors were revised or strengthened so as to insure their usefulness. In a few cases where effective foils could not be devised another item was used in its place.

2. Evidence of Reliability

The distribution of scores for each test were examined for central tendency (42:24) and standard deviation (42:57). The odd-even reliability for each test was then estimated with the use of the Spearman-Brown prophecy formula (42:332). The formula is:

$$r_{xx} = \frac{2 r_{oe}}{1 + r_{oe}}$$

where r_{xx} is the coefficient of reliability of the test and r_{oe} is the coefficient of correlation between odd and even items. In Table I a summary of these data for each test is provided.

Inspection of Table I reveals relatively lower reliability for the knowledge test. The small number of items together with their restricted variation probably accounts for the somewhat low coefficients. These results on the trial test made it evident that the number of items could not be reduced greatly or the reliability of the tests would be low.

TABLE I

Means, Standard Deviations and Spearman-Brown
Estimates of Reliability for Tests

| Test | No. of Items | Mean | Standard Deviation | Odd-Even Correlation | Reliability Coefficient |
|----------------|-----------------|-------|-----------------------|-------------------------|----------------------------|
| Knowledge | 24 | 16.16 | 3.48 | .483 | .651 |
| Translation | 30 | 16.20 | 4.43 | .564 | .721 |
| Interpretation | 24 | 12.48 | 3.07 | .516 | .681 |
| Extrapolation | 26 | 13.48 | 3.02 | .542 | .703 |

3. Homogeneity of Test Behavior

To ascertain whether or not the four types of tests constructed were measuring separate behaviors, an F-test of significance for departure from homogeneity was applied. This technique proposed by Heidt is designed to determine whether individuals react in a significantly different manner to items between or among areas in a test than they do to items within areas. The technique does not guarantee homogeneity within a given area but indicates a relatively greater lack of homogeneity between or among areas than within areas. The values of F are determined by the formula

$$F = \frac{1 + \bar{R}_V - 2 \bar{R}_A}{1 - \bar{R}_V}$$

where \bar{R}_V is the average intra-area coefficient and \bar{R}_A is the average inter-area coefficient of correlation.

Scores were obtained separately for the odd and even items of each test. Inter- and intra-area correlations necessary for substitution into the formula are shown in Table 2.

TABLE 2

Intra- and Inter-Correlations Between Tests

| Test | | Test | | | | | | | |
|----------------|------|-----------|------|------------------|------|---------------------|------|---------------|------|
| | | Knowledge | | Transla- tion | | Interpreta- tion | | Extrapolation | |
| | | Odd | Even | Odd | Even | Odd | Even | Odd | Even |
| Knowledge | Odd | | | .576 | .578 | .536 | .282 | .389 | .458 |
| | Even | .483 | | .422 | .288 | .513 | .480 | .361 | .498 |
| Translation | Odd | | | | | .419 | .317 | .344 | .476 |
| | Even | | | .564 | | .533 | .250 | .324 | .483 |
| Interpretation | Odd | | | | | | | .426 | .607 |
| | Even | | | | | .516 | | .223 | .354 |
| Extrapolation | Odd | | | | | | | | |
| | Even | | | | | | | .542 | |

The correlations were averaged according to the function $\frac{1}{2} \log_e \frac{1+r}{1-r}$, the resulting intra- and inter-area correlations substituted into the formula and F-values obtained for each pair of tests. These values are reported in Table 3.

Inspection of the data in Table 3 shows that the values of F between translation and interpretation, translation and extrapolation and interpretation and extrapolation are significant at the five percent level of confidence indicating that these tests are heterogeneous with respect to each other. The knowledge test does not depart significantly from the others, although the F-values closely approach the five percent level. There seemed to be sufficient evidence to warrant the continued use of the

TABLE 3

Values of F Between Tests For
Departure From Homogeneity

| Test | Test | | |
|----------------|-------------|----------------|---------------|
| | Translation | Interpretation | Extrapolation |
| Knowledge | 1.21 | 1.17 | 1.30 |
| Translation | | 1.68* | 1.64* |
| Interpretation | | | 1.49* |

Required for significance, 74 and 74 degrees of freedom, 1% = 1.74**
5% = 1.47*

four separate tests in light of the fact that a further check on the heterogeneity of the different tests would be made with a larger sample at the time of second testing in the retention experiment.

D. Administration of the Refined Test

1. Description of the Test

The refined test contained ninety-five test items and required approximately one and one-half to two hours for administration. The items were distributed according to behavior measured as follows:

| <u>Behavior</u> | <u>Number of Items</u> |
|-----------------|------------------------|
| Knowledge | 24 |
| Translation | 24 |
| Interpretation | 23 |
| Extrapolation | 24 |
| Total | 95 |

These items were combined in a single test booklet entitled "Examination on Tests and Measurements". Although the knowledge items appear first in the booklet, the remainder are somewhat scattered, as often questions measuring a different behavior refer to the same group of data. The

booklet is prefaced with a statement of the purpose of the examination for the testee's benefit. A copy is included in Appendix C.

From the original trial test twenty-nine items were eliminated and four entirely new items were added. Fourteen of the trial test items were eliminated on the basis of low or negative item-criterion correlations. Seventeen items, upon further examination of responses and individual distractors, were found to be weak items and could not be effectively revised to eliminate ambiguity and to insure effectiveness of all distractors. Four new items were added and were patterned after items that seemed to be statistically and logically valid items.

2. Description of the Subjects

The test was administered as a pretest to 306 students enrolled in Education 62. At the time of the final examination 310 students took the test. The scores of this latter group were used for analysis of the examination.

The majority of these students were sophomores comprising approximately eighty percent of the group. The remainder were juniors and seniors, with a very small number of seniors in the sample. Elementary education was the most frequently listed major, but the sample included students majoring in twenty-nine different fields. The students had all taken Education 61, or its equivalent, as a prerequisite course to Education 62. For those listed, this requirement was satisfied by Psychology 70 for twenty students, Home Economics 91 for nine students and approximately ten percent by psychology credit from another school.

The Linguistic or L-score on the American Council on Education Psychological Examination was obtained for 255 of the sample students.

These scores were obtained for the purpose of correlating a measure of scholastic aptitude with each separate test.

Of the 310 subjects, 301 had also taken the pretest. These 301 students were requested approximately one semester later to participate in a retest. Some of these students were requested by instructors in various classes and some by mail. A copy of the request letter is shown in Appendix D. The results of 172 students obtained approximately four months later were examined in the study of retention.

E. Treatment of the Data

1. Item Difficulty and Discrimination

Each of the test items was then analyzed to determine item difficulty and the correlation of each item with its respective test total. The difficulty was expressed in terms of percentage of the group who marked it correctly. The discrimination ability of each item was estimated with the use of an item analysis table. (10)

2. Reliability of the Tests

The reliability of each test was estimated using the Spearman-Brown and Kuder-Richardson techniques. The Spearman-Brown formula was shown earlier in this chapter. The Kuder-Richardson formula is: (42:334)

$$r_{KR} = \left(\frac{n}{n-1} \right) \left(\frac{\sigma_t^2 - p^2}{\sigma_t^2} \right)$$

where n = the number of items in the test

p = the proportion of subjects responding correctly to an item

$q = 1-p$

σ_t^2 = the variance of the total test scores

Both of these formulas assume homogeneity of test content and will yield similar results if this assumption is met.

3. Homogeneity of Tests

The F-test for departure from homogeneity was applied to the results of this test. This analysis was for the purpose of further investigating the feasibility of using the separate test results in the retention study. The same test was previously applied to the trial test results and has been described.

4. Evidence of Validity

Though logical validity of the examination questions was established by the relating of the questions to the course objective, further evidence of validity was obtained using a semi-external criterion. The final course marks were correlated with the results of each test. In order to take into account the differing degrees of variation among the grading practices of the various instructors the final grades were converted to standard units. To do this, the means and standard deviations were computed for the grade distribution of each instructor, the separate grades converted to a standard score and then each standard score was converted to a scaled unit employing a ten-unit scale.

The final mark included the results of the test used in the experiment and therefore the results are somewhat spurious. However, inasmuch as the course grades represent the combined results of all types of appraisal by the instructor, it would seem to furnish important evidence concerning the appropriateness of the test. The weight given to this test as a part of the course grade varied among instructors, but in most cases comprised less than one-sixth of the final mark.

Correlation coefficients were computed between the American Council on Education L-Scores and each of the tests. The L-scores were expressed in stanines, a one to nine standard scale, and these scores were available on 255 students who took the test.

5. Retention

Of the 301 students who took both the pretest and test after completion of the unit, 172 responded to the request to take the retest. It was necessary to determine whether or not this group was a suitable sample from the parent population. The hypothesis that there was no significant difference between the group that took the retest and the group that did not was evaluated using the formula for separate group variance which is: (42:130)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{x_1^2}{K_1(K_1-1)} + \frac{x_2^2}{K_2(K_2-1)}}$$

where \bar{x}_1 is the mean of the first group

\bar{x}_2 is the mean of the second group

x_1^2 is the sum of the squared deviations of the scores away from the mean of the first group

x_2^2 is the sum of the squared deviations of the scores away from the mean in the second group

K_1 is the number of cases in the first group

K_2 is the number of cases in the second group

The obtained value of t is evaluated for significance using (K_1-1) or (K_2-1) degrees of freedom.

To discover the degree of relationship between the scores on each test administration, correlation coefficients were computed between the

pretest and the test, the test and the retest, and the pretest and the retest. The correlations reveal the degree to which changes in one of the variables are accompanied by changes in the other variable.

Further, to determine if the differences between the means of scores on each of these test administrations were significant beyond chance, a test of significance for correlated data was employed. To do this it was first necessary to find the standard error of differences between the means. The formula for this is as follows: (17:186)

$$\sigma_{dM} = \sqrt{\sigma^2_{M_1} + \sigma^2_{M_2} - 2 r_{12} \sigma_{M_1} \sigma_{M_2}}$$

where $\sigma^2_{M_1}$ = Standard error of the mean of the first distribution

$\sigma^2_{M_2}$ = Standard error of the mean of the second distribution

r_{12} = Correlation between the two sets of means

The standard error of the mean may be estimated directly from the sum of squares using the formula: (17:65)

$$\sigma_M = \sqrt{\frac{\sum x^2}{N(N-1)}}$$

where $\sum x^2$ = the sum of the squared deviations from the mean

N = the number of cases in the sample

The standard error of the differences between the means may then be divided into the differences between the means to obtain a t-value. This t-value is evaluated using the t-table with $N-1$ degrees of freedom.

Retention was also studied by computing the average gain retained for each of the tests. To do this the differences between the pretest and test were summated, and divided into the sum of the differences between pretest and retest. The results were expressed in terms of the percent of the material learned in the course that was retained approximately four months later.

To ascertain whether or not these percentages were significantly different from each other it was necessary to employ tests of significance of difference between percentages. The formula for the standard error of the difference between correlated percentages is: (17:19)

$$\sigma_{d_p} = \sqrt{\sigma_{P_1}^2 + \sigma_{P_2}^2 - 2 r_{12} \sigma_{P_1} \sigma_{P_2}}$$

where σ_{P_1} = standard error of the first percentage

σ_{P_2} = standard error of the second percentage

r_{12} = correlation of percentages in pairs of samples

The standard error of a percentage may be found by the formula: (17:175)

$$\sigma_p = \sqrt{\frac{pq}{N}}$$

where p = percentage in the selected category

$q = 1-p$

N = Number in the sample

The obtained standard error of the difference may then be divided into the difference between the percentages to obtain a t-value. This value may be evaluated using $N-1$ degrees of freedom.

To discover further evidence of retention, the individual items of all tests were studied. The following item combinations were tabulated:

- 1) Wrong on pretest - right on the test - wrong on retest.
- 2) Wrong on pretest - right on the test - right on retest.

It was assumed that items which frequently elicited the first pattern measured material that was learned in the course but was likely to be forgotten. The second pattern would indicate items that gave evidence of permanence of behavior learned. Since the items were all multiple-choice, the factor of chance for each of these combinations was subtracted before frequencies were considered as evidence. The individual items were then

examined to note the type of behavior they measured or the occurrence of any types of item patterns other than those originally set forth in the construction of the tests.

CHAPTER III

FINDINGS OF THE STUDYA. Analysis of the Tests

The analysis of the tests will be considered first. The results of the retention study will be reported in the latter part of the chapter.

1. Item Difficulty

The test item difficulty reported in terms of the percent of the group who responded correctly to each item is shown in Table 4. For the knowledge test the average level of difficulty was 62.13 percent. The averages for translation, interpretation and extrapolation were 60.45, 60.61 and 56.52, respectively. Examination of the individual difficulty percentages shows that the items tend to cluster about the means with slightly more items in the upper percentages. The percentages on all tests seem to be well distributed with no items either being answered correctly or missed by one hundred percent of the group. Few items on any tests were below the twenty percent level of difficulty, but a few were above the ninety percent level, ninety-four being the highest.

Although it can be shown mathematically that maximum discrimination can be achieved when the average difficulty is fifty percent, such percentages as are reported are distributed adequately to achieve sufficiently high discrimination. Furthermore, a test which has an average difficulty higher than fifty percent is less likely to be discouraging to the testee.

2. Item Discrimination

The correlation coefficients for each item with its respective test total are shown in Table 4. The correlations are also reported for each item with the total score when all tests are combined. These latter correlations will be analyzed in the following section.

The first step in evaluating these correlations was to compute the standard error of a correlation coefficient for this sample. The formula for the standard error became $S.E. = \frac{1}{\sqrt{310-1}}$ and the resulting value was found to be .057. Only two such items were found to exceed this limit, both being in the translation test, numbers 31 and 35. These items were eliminated in that such correlations would indicate that they were not effectively performing the same function as the other items in the same test. These two items, being eliminated, were not considered further or used in the retention test.

Most of the item correlations were sufficiently high to be considered highly discriminating. If a coefficient of .40 were arbitrarily selected as a highly desirable standard, fifty such items appear in the test, or approximately fifty-four percent of the total items. The items in the knowledge and translation tests contributed fewer highly discriminating items than did the other tests. In the knowledge test, thirty-seven percent of the items were above this standard and forty-two percent were above the standard on the translation test. For the interpretation and extrapolation tests the corresponding percents were sixty-five and sixty-three, respectively.

TABLE 4

Item Difficulty and Item Discrimination
for Test Score and Combined Total Score

| Test | Item Number | Percent Correct Response | Correlation with Test Score | Correlation with Combined Total Score |
|-------------|-------------|--------------------------|-----------------------------|---------------------------------------|
| Knowledge | 1 | 75 | .60 | .36 |
| | 2 | 74 | .22 | .14 |
| | 3 | 58 | .36 | .21 |
| | 4 | 95 | .22 | .34 |
| | 5 | 55 | .26 | .17 |
| | 6 | 76 | .36 | .31 |
| | 7 | 39 | .31 | .12 |
| | 8 | 52 | .26 | .13 |
| | 9 | 45 | .41 | .32 |
| | 10 | 74 | .36 | .25 |
| | 11 | 56 | .39 | .35 |
| | 12 | 65 | .67 | .50 |
| | 13 | 40 | .44 | .27 |
| | 14 | 36 | .42 | .27 |
| | 15 | 92 | .53 | .16 |
| | 16 | 51 | .37 | .27 |
| | 17 | 86 | .17 | .03 |
| | 18 | 62 | .15 | .10 |
| | 19 | 78 | .56 | .36 |
| | 20 | 76 | .51 | .46 |
| | 21 | 21 | .24 | .25 |
| | 22 | 74 | .54 | .58 |
| | 23 | 85 | .32 | .16 |
| | 24 | 34 | .31 | .10 |
| Translation | 25 | 45 | .52 | .31 |
| | 26 | 59 | .58 | .43 |
| | 27 | 73 | .40 | .33 |
| | 28 | 91 | .28 | .21 |
| | 29 | 93 | .50 | .45 |
| | 30 | 28 | .21 | .20 |
| | 31 | 65 | .00 | -.15 |
| | 32 | 64 | .38 | .26 |
| | 33 | 13 | .55 | .30 |
| | 34 | 5 | .40 | .34 |
| | 35 | 67 | -.27 | .00 |
| | 36 | 35 | .28 | .19 |
| | 37 | 72 | .33 | .37 |
| | 38 | 60 | .30 | .20 |
| | 39 | 69 | .45 | .34 |
| | 40 | 60 | .41 | .34 |

Table 4 (Continued)

| Test | Item Number | Percent Correct Response | Correlation with Test Score | Correlation with Combined Total Score |
|------------------------|-------------|--------------------------|-----------------------------|---------------------------------------|
| Translation (Cont.) | 41 | 68 | .37 | .30 |
| | 54 | 85 | .20 | .07 |
| | 63 | 77 | .36 | .42 |
| | 66 | 52 | .38 | .12 |
| | 82 | 93 | .37 | .23 |
| | 83 | 68 | .43 | .48 |
| | 86 | 54 | .45 | .38 |
| | 89 | 66 | .47 | .54 |
| Interpretation | 92 | 90 | .49 | .32 |
| | 43 | 44 | .39 | .22 |
| | 44 | 88 | .63 | .41 |
| | 45 | 50 | .42 | .24 |
| | 46 | 45 | .45 | .32 |
| | 47 | 93 | .39 | .46 |
| | 53 | 94 | .34 | .51 |
| | 55 | 89 | .51 | .50 |
| | 57 | 76 | .55 | .38 |
| | 58 | 13 | .37 | .36 |
| | 59 | 92 | .52 | .26 |
| | 60 | 40 | .46 | .38 |
| | 64 | 69 | .45 | .50 |
| | 67 | 30 | .39 | .37 |
| | 69 | 39 | .48 | .54 |
| | 72 | 58 | .32 | .26 |
| | 73 | 57 | .51 | .59 |
| | 75 | 58 | .32 | .25 |
| | 84 | 82 | .53 | .39 |
| | 85 | 81 | .54 | .25 |
| 92 | 42 | .66 | .56 | |
| 93 | 34 | .66 | .51 | |
| 94 | 22 | .35 | .19 | |
| Extrapolation | 48 | 64 | .44 | .47 |
| | 49 | 94 | .32 | .34 |
| | 50 | 19 | .54 | .48 |
| | 51 | 74 | .44 | .44 |
| | 52 | 91 | .43 | .37 |
| | 56 | 63 | .17 | .00 |
| | 61 | 42 | .32 | .08 |
| | 62 | 54 | .45 | .51 |
| | 65 | 23 | .45 | .24 |
| | 68 | 64 | .33 | .26 |
| | 70 | 55 | .38 | .36 |
| | 71 | 55 | .26 | .15 |
| | 74 | 41 | .35 | .19 |
| | 76 | 56 | .50 | .42 |
| | 77 | 67 | .38 | .40 |
| 78 | 34 | .52 | .42 | |

Table 4 (Concluded)

| Test | Item Number | Percent Correct Response | Correlation with Test Score | Correlation with Combined Total Score |
|--------------------------|-------------|--------------------------|-----------------------------|---------------------------------------|
| Extrapolation (Cont.) | 79 | 33 | .45 | .28 |
| | 80 | 92 | .53 | .53 |
| | 81 | 39 | .46 | .36 |
| | 87 | 91 | .56 | .37 |
| | 88 | 76 | .28 | .14 |
| | 90 | 33 | .85 | .73 |
| | 91 | 50 | .58 | .58 |
| | 95 | 44 | .49 | .51 |

Setting of such arbitrary standards for appraisal of test item correlations as a rule depends on the use made of the test. When selection of items is being made it is appropriate to choose first the items yielding the highest correlations and to accept the lower ones as necessary. With regard to pre-set standards in the case of a test, Lindquist (22:315) suggests that correlations of .20 to .30 are characteristic of heterogeneous test material and somewhat higher coefficients characterize more homogeneous materials. By construction, the tests in this study were intended to measure somewhat homogeneous behaviors and, therefore, .30 might be accepted as a desirable lower limit to appraise the items. Only fifteen, or seventeen percent of the items of the total group, fall below this limit. Most of the items are in the knowledge and translation tests with six in each. The remaining three are in the test on extrapolation.

3. Evidence of Reliability

The Spearman-Brown and the Kuder-Richardson estimates of reliability are shown in Table 5. The relatively small number of items included in each test is the major reason for the somewhat low reliabilities reported.

TABLE 5

Spearman-Brown and Kuder-Richardson
Estimates of Reliability

| Test | Number of Items | Odd-Even Correlation | Spearman- Brown Estimate | Kuder- Richardson Estimate |
|----------------|-----------------------|-------------------------|--------------------------------|----------------------------------|
| Knowledge | 24 | .297 | .458 | .495 |
| Translation | 22 | .290 | .450 | .507 |
| Interpretation | 23 | .477 | .646 | .531 |
| Extrapolation | 24 | .440 | .611 | .537 |

In answer to the question of how reliable a test must be in order to meet standards of acceptability, Kolley (22:609) has been widely quoted. He suggests that a minimum reliability coefficient of .50 be set to evaluate level of group accomplishment. Using such a standard, the tests employed in this experiment meet an acceptable level when referring to the Kuder-Richardson estimates. The knowledge and translation tests fall slightly below using the Spearman-Brown estimates.

Certainly higher reliability would be more desirable for tests having such functions to perform as in this experiment. However, the limiting factor of testing time makes it tremendously difficult to employ tests with larger number of items. To do so would necessitate the elimination of some of the tests included.

4. Homogeneity of Tests

One positive indication of homogeneity of the behavior measured by the different tests can be noted by inspection of the item correlations

in Table 4. Since the total score constitutes the criterion with which each item is compared, the higher the correlation the more the behavior of each item is like the behavior measured by the total test. It may be noted that when all of the tests are combined into one single test score and the items correlated with this total, most of the item coefficients are reduced. This reduction would indicate a greater heterogeneity of test content when the tests were combined or conversely a greater homogeneity of content in the separate tests.

Closer inspection of the individual item correlations reveals that few of the items on any test increase in discrimination ability when the total test is used as a criterion. There are four such items in the knowledge test, six in the test of translation, five in interpretation and five in the extrapolation test. The differences are, in most cases, very small, however, and all of the rest of the items show higher correlations when using their own respective tests as the criterion. It can be concluded from this analysis that there is greater homogeneity of behavior within the tests than there is when the tests are combined. It is not possible from Table 4, however, to determine the degree to which each test is homogeneous with respect to each other test. To test this hypothesis an F-test for departure from homogeneity was applied. Intra- and inter-area correlations necessary for substitution into the formula are shown in Table 6.

For each test the inter- and intra-area correlations were averaged and the F-values computed. The resulting F-values are shown in Table 7.

Inspection of Table 7 shows that, with 309 and 309 degrees of freedom, the resulting F-values are significant beyond the one percent level of confidence between knowledge and translation, knowledge and interpretation

TABLE 6

Intra- and Inter-Area Correlations Between Tests

| Subtest | | Subtest | | | | | | | |
|----------------|------|-----------|------|------------------|------|---------------------|------|--------------------|-------|
| | | Knowledge | | Transla- tion | | Interpre- tation | | Extrapola- tion | |
| | | Odd | Even | Odd | Even | Odd | Even | Odd | Even |
| Knowledge | Odd | | | -.139 | .109 | .266 | .253 | .272 | .284 |
| | Even | .297 | | .291 | .347 | .287 | .351 | .323 | .432 |
| Translation | Odd | | | | | .330 | .209 | .387 | -.056 |
| | Even | | | .290 | | .316 | .313 | .288 | .292 |
| Interpretation | Odd | | | | | | | .491 | .537 |
| | Even | | | | | .477 | | .311 | .399 |
| Extrapolation | Odd | | | | | | | | |
| | Even | | | | | | | .440 | |

TABLE 7

Values of F For Tests of Homogeneity
Between Tests

| Subtest | Translation | Interpretation | Extrapolation |
|----------------|-------------|----------------|---------------|
| Knowledge | 1.358** | 1.332** | 1.176 |
| Translation | | 1.310* | 1.423** |
| Interpretation | | | 1.085 |

Required for significance, 309 and 309 degrees of freedom, 1% = 1.22**
5% = 1.32*

and translation and extrapolation. The F-value for translation and interpretation is significant at the five percent level. The F-values between knowledge and interpretation and interpretation and extrapolation were not significant. The hypothesis that behaviors measured by the different tests were heterogeneous with respect to each other can then be retained between all tests except knowledge and extrapolation and interpretation and extrapolation. It may also be noted that even though the F-value between knowledge and extrapolation did not reach the five percent level of confidence it closely approached this value.

It would seem desirable on the basis of this analysis to combine the interpretation and extrapolation tests since they do not perform separate functions. This was done, and the resulting F-values when these two tests were combined are shown in Table 8.

TABLE 8

Values of F for Tests of Homogeneity Between Tests
(Interpretation - Extrapolation Combined)

| Test | Translation | Interpretation- Extrapolation |
|--|-------------|----------------------------------|
| Knowledge | 1.388** | 1.358** |
| Translation | | 1.489** |
| Required for significance, 618 and 309 degrees of freedom, 1% = 1.22** 5% = 1.32* | | |

Inspection of Table 8 with 618 and 309 degrees of freedom, shows that all values of F are significant beyond the one percent level of confidence. The hypothesis that these three tests are heterogeneous with respect to each other can be retained.

As a result of this analysis the tests for interpretation and extrapolation were combined for the retention study described in the latter part of this chapter.

5. Evidence of Validity

The logical or content validity of the examination would seem to be fairly clearly established through the process of carefully matching the content of each item with the definition as set forth in the Taxonomy. Further assurance would be obtained by carefully analyzing each test item to eliminate ambiguity of the question or any distractors employed.

A semi-external criterion, namely final course marks, was also employed to obtain a measure of empirical validity. The resulting correlations between the tests and the final grades were as follows:

| <u>Test</u> | <u>r</u> |
|----------------------------------|----------|
| Knowledge | .623 |
| Translation | .567 |
| Interpretation- Extrapolation | .525 |

The final course mark for Education 62 included the results of these tests as a part of the course grade. The correlations reported are somewhat spurious, although for most instructors the test constituted little more than one-sixth of the final grade.

Assuming that the combined methods of appraisal, i.e. the final course marks, represent valid estimates of the degree to which the intended objectives are accomplished, a high positive relationship between the tests and the grades would be expected. Such a relationship is shown, although the reported coefficients might be expected to be somewhat higher. The slightly higher coefficient between the knowledge test and

the course mark perhaps suggests that the final course mark contains a more generous allotment of types of appraisal measuring this objective.

It seemed also highly desirable that the relationship between each test and a measure of scholastic aptitude be found. The L-score of the American Council on Education Psychological Examination was used for this purpose and the resulting correlation coefficients between this value and each test are as follows:

| <u>Test</u> | <u>r</u> |
|----------------------------------|----------|
| Knowledge | .364 |
| Translation | .362 |
| Interpretation- Extrapolation | .343 |

It is evident from the inspection of these coefficients that the magnitude of relationship between the L-score and each test is very similar. It would seem safe to assume that the influence of the scholastic aptitude factor, as measured by the American Council on Education Psychological Examination, is equally present in the performance required by the separate tests.

B. The Study of Retention

In the previous section evidence was found indicating that two of the original four tests should be combined, in that they did not seem to be performing separate functions. Further evidence indicated that once two of these tests were combined the resulting three tests measured functions heterogeneous with respect to each other. The three tests, namely: knowledge; translation and interpretation-extrapolation; were used in the study of retention. The test was readministered approximately

four months after the test at the end of the unit to 172 of the original students who took both the pretest and the unit test. In the following pages the nature of the material retained is analyzed.

1. The Retention Group

The students who took the retest were contacted both through other classes and by mail. The largest percentage in any single course from which test results were obtained was found in Education 141 where approximately twenty-six percent were contacted. Several students were contacted in elementary education courses and a few in home economics. The remainder were scattered and were reached directly by mail or contacted personally. A small group did report and took the test as a group according to a prearranged schedule. The instructions accompanying the administration are essentially contained in the letter, a copy of which is in Appendix D.

The possibility that subject matter learned in other courses might transfer was considered. Two courses likely to yield large amounts of positive transfer in the area being tested were courses in Educational Measurement and Evaluation, Education 263, and Principles and Practices of Guidance, Education 269. Upon inquiry it was found that these courses were not offered during the semester corresponding with the retention period. Inquiry and examination of the content of Education 141, Principles and Practices of Teaching in the Secondary Schools, revealed that very little was covered in the area of tests or testing during the semester involved. Apparently large amounts of transfer would not be derived from the courses the students were taking during the retention period.

To determine whether or not the sample who took the retest was characteristic of the population from which it was drawn, a significance test was applied. The test scores of the students at the end of the unit were utilized. The hypothesis was tested that there was no difference between the group who took the retest and the group who did not. The means and t-statistics are reported for the difference between the sample and non-sample for each test.

| <u>Test</u> | <u>Mean of Retest Group</u> | <u>Mean of Non-Retest Group</u> | <u>t</u> |
|----------------------------------|-----------------------------|---------------------------------|----------|
| Knowledge | 15.24 | 14.60 | 1.72 |
| Translation | 13.83 | 12.92 | 2.68** |
| Interpretation- Extrapolation | 27.88 | 27.47 | .58 |

For 171 degrees of freedom a value of 1.96 is required at the five percent level and 2.58 at the one percent level of confidence. The hypothesis that there is no difference in groups can be retained for the knowledge and interpretation-extrapolations tests and rejected in the case of the test of translation. However, in light of the fact that 172, or fifty-seven percent, of the original population was obtained in the sample and there was only one test that showed a significant difference, it was concluded that the group retested was representative of the total group.

2. Relationships Between Pretest - Test - Retest

The resulting correlation coefficients between the three administrations of each test are shown in Table 9.

It can be noted by inspection of these correlations that the relationship of performance on the successive administration is positive in

TABLE 9

Correlation Coefficients Between the Three
Administrations of the Tests

| Test | Correlation Coefficients | | |
|----------------------------------|--------------------------|---------------|------------------|
| | Pretest - Test | Test - Retest | Pretest - Retest |
| Knowledge | .391 | .409 | .274 |
| Translation | .492 | .289 | .470 |
| Interpretation- Extrapolation | .537 | .599 | .489 |

all cases, although not to a high degree. Slightly higher relationships reported for the interpretation-extrapolation test suggest that relative performance on this test is more nearly the same than on the knowledge and translation tests. These correlations are not directly comparable, however, due to the smaller number of items and lower reliabilities of the knowledge and translation tests.

Such correlations indicate the extent to which individuals tend to maintain their relative rank on the tests on successive administrations. They do not reveal the magnitude of difference in performance from test to test.

3. Differences in Average Performance for the Various Test Administrations

To determine if the difference in mean performance on the various test administrations were significant, a t-test for correlated data was applied. In Table 10 the differences together with the accompanying t-values are shown.

TABLE 10

Differences in Mean Scores and the Accompanying t-value
Between the Three Administrations of the Tests

| Test | Combination | | |
|----------------------------------|---------------------------|--------------------------|--------------------------|
| | Pretest - Test | Test - Retest | Pretest - Retest |
| Knowledge | Diff = -3.65 t = 10.42 | Diff = 2.65 t = 10.19 | Diff = -1.00 t = 3.33 |
| Translation | Diff = -2.78 t = 13.29 | Diff = 2.04 t = 8.16 | Diff = -.74 t = 2.74 |
| Interpretation- Extrapolation | Diff = -5.18 t = 12.33 | Diff = 4.10 t = 9.11 | Diff = -1.08 t = 2.57 |

Required for significance, 171 degrees of freedom, 2% = 2.32
1% = 2.58

It is revealed by inspection of these t-values that all differences are significant beyond the one percent level of confidence except the difference between the pretest and retest for interpretation-extrapolation. This value is significant at the two percent level. These results show that, on the average, a significant amount of material was learned in the course, a significant amount was forgotten during the retention period, and at the end of the retention period the students still retained enough learning so that their performance was significantly different from that at the time of the pretest.

4. The Retention Graph

Figure 1 makes possible the comparison of the course of learning and retention of the materials on the three tests. The mean performances are reported in percent of items correct and therefore may be directly compared.

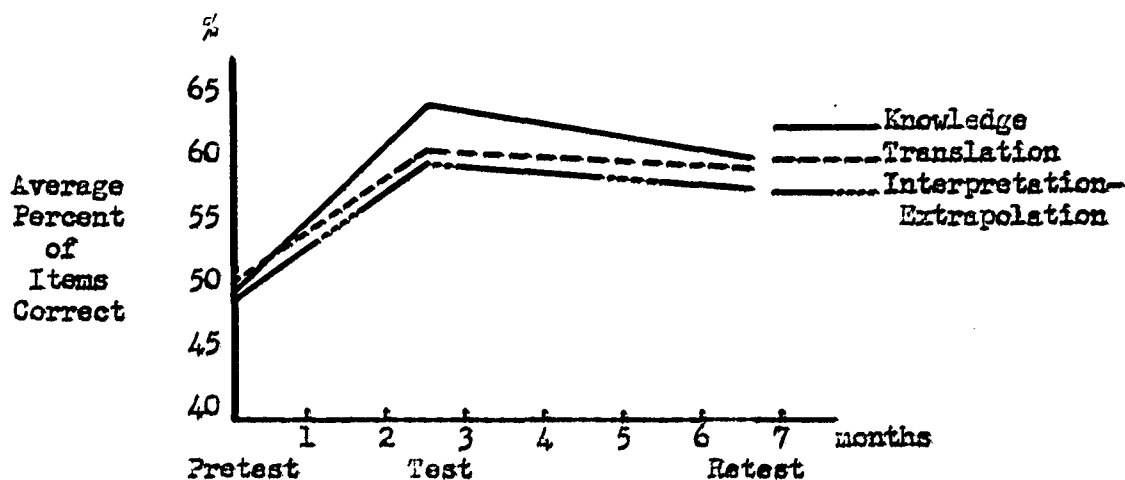


FIGURE 1

Retention Graph of the Three Tests: Knowledge, Translation,
Interpretation-Extrapolation

Examination of Figure 1 indicates the greatest gain was made on the knowledge test in terms of average percent of items correct. The averages increased from 49.6 percent to 64.8 percent. It may also be noted that knowledge performance dropped relatively more than the other tests with an average of 60.6 percent of the items correct on the retest. The average for translation began with 50.2 percent correct, increased to 62.4 percent and then dropped to 59.5 percent. The corresponding percents for interpretation-extrapolation are 49.2, 60.2 and 57.9 percent.

5. Average Percent of Gain Retained

The amount of material retained for each test may also be expressed in terms of percent of gain retained. These percentages are as follows:

| <u>Test</u> | <u>Percent of Gain Retained</u> |
|----------------------------------|---------------------------------|
| Knowledge | 71.32 |
| Translation | 73.96 |
| Interpretation- Extrapolation | 63.57 |

Examination of these percentages shows that a large amount of retention was demonstrated on all three tests. After four months the students remembered approximately three-fourths of the knowledge they learned about tests and measurement and the ability to translate this knowledge. The abilities and skills they learned in being able to interpret and extrapolate from data showed even more permanence with a reported percentage of 63.57.

To find if the differences between these percentages were significant beyond chance a test of significance of difference between percentages was utilized. The differences between these percentages with the accompanying t-values are as follows:

| <u>Tests</u> | <u>Differences Between Percentages</u> | <u>t</u> |
|--|--|----------|
| Knowledge and Translation | 2.64 | .629 |
| Knowledge and Interpretation- Extrapolation | 12.3 | 3.001** |
| Translation and Interpretation- Extrapolation | 9.1 | 2.597** |

Consulting a t-table with N-1 or 171 degrees of freedom, it is shown that the differences between knowledge and interpretation-extrapolation and between translation and interpretation-extrapolation are significant beyond the one percent level of confidence and the difference between knowledge and translation is not significant at the five percent level.

It may be concluded that a significantly larger percentage of gain made in the course is retained on the interpretation-extrapolation test than on the knowledge and translation tests. These results suggest that such behaviors as interpretation and extrapolation demonstrate a higher degree of permanency than do recall of knowledge and translation.

6. Occurrence of Certain Patterns on the Successive Administrations of the Test Items

In hope of finding further evidence of retention, two different patterns of responses to the individual items were examined. These patterns were: wrong on the pretest, right on the test, and wrong on the retest; and, wrong on the pretest, right on the test, and right on the retest. It was felt that the first pattern would reveal correct responses learned during the course and forgotten by the time of the retest. The second pattern would show responses learned during the course and retained at the time of the retest. The frequency of such patterns is shown in Appendix E.

Because the items on the tests were multiple-choice type, containing either four or five choices for each question, it was necessary to adjust the occurrence of each pattern to allow for chance. The wrong-right-wrong combination could be achieved nine times in sixty-four by chance if the item were four-choice and sixteen times in 125 if the item contained five choices. The wrong-right-right pattern would occur by chance three out of sixty-four times for the four choice items and four in 125 times for the five-choice items. Using these ratios, each item was then adjusted by subtracting chance occurrences from the tabulated frequency. Only those responses in excess of chance were considered for further analysis. The adjusted frequencies are also shown in Appendix E.

Inspection of the frequencies reveals that there were sixteen items remaining in the wrong-right-wrong category after chance was subtracted. Apparently correct answers were learned to these items during the study of the units but were forgotten more often than could be accounted for by chance. When these items were matched with the objectives being considered in this study it was found that four items fell into each of the four tests; knowledge, translation, interpretation and extrapolation.

Examination of wrong-right-right frequencies indicates that eighty-nine items apparently measured behaviors retained beyond chance occurrence. Again, as in the case of the previous patterns, no particular test contained a preponderance of these items.

Four of the ninety-three items on the test demonstrated neither forgetting nor retention of what was learned in the course. Apparently only chance responses to the items followed a wrong-right pattern on the first two tests.

The general conclusion that can be drawn from this type of analysis is that the behaviors measured by all of the tests demonstrate a higher degree of retention than forgetting. It is impossible to conclude from the item data whether or not any particular behavioral outcomes demonstrate different degrees of permanency from others.

CHAPTER IV

IMPLICATIONS OF THE INVESTIGATION

1. The results of this study indicate the need for careful delineation of course objectives. Such objectives need to be defined in terms of student behavior as well as to contain comprehensive enumeration of the different aspects of course content. The homogeneity study in this experiment showed that tests constructed to measure certain behavioral outcomes apparently perform separate functions as evaluation devices. Thus, to insure that multiple course outcomes, in line with the objectives of instruction, are achieved, it becomes necessary to design evaluation instruments to accomplish these separate functions. As was concluded in the review of literature in Chapter I, the achievement of one objective cannot be inferred from the measured attainment of another.

This conclusion is also given weight by the result of the study of retention. These results indicated that the degree of permanency of different course outcomes is relatively higher for some behaviors than for others. Therefore, in order to appraise the results and effectiveness of instructional programs it is necessary to have a thorough knowledge of the type and nature of objectives being emphasized.

2. The usefulness of the "Taxonomy of Educational Objectives" as a classification with which to guide the construction of evaluation instruments seems to be borne out. Previously, Urdal (37) verified at least four major categories of the Taxonomy using factor analysis. The categories were knowledge, comprehension, application and analysis. The results of the analysis of student behaviors in the present study further suggest

that at least one of the major categories in the Taxonomy, i.e. comprehension, contains types of behavioral outcomes that should be considered separately. It was found that the abilities and skills involved in translation involved a separate evaluative function from interpretation and extrapolation which fall in this same major category. These latter two, however, involved similar functions.

3. In that the organization of the Taxonomy is hierarchial in nature, i.e. each higher level of intellectual ability is built on and includes the lower levels, it is suggested by this study that greater emphasis be given to the higher level outcomes. It was shown that interpretation-extrapolation demonstrated higher degrees of retention. Since the ability to interpret and extrapolate presumes the inclusion of knowledge and the ability to translate knowledge, it seems evident that sound instructional technique would emphasize these objectives of more lasting value. The alternative would be to adopt the less economical practice of placing major emphasis on the lower level outcomes. The latter has often been the practice. Tyler (21) found that interviews with college students indicated that more than sixty percent of the students in college believe their chief duty is to memorize information. Tyler stated that the emphasis given to recall of fact in the typical college examination is one of the chief reasons for the existence of this belief.

4. The majority of studies, some of which were discussed in Chapter I, suggest that much of what is learned in school is forgotten. This fact has long been the concern of educators. From this standpoint, the results of the present investigation seem encouraging. To place greater emphasis on these more permanent yet inclusive types of behavioral objectives seems

to provide at least a partial solution to the problem. Such a practice should result in optimum learning experiences due both to the inclusive nature of the higher level objectives and to their greater permanency.

5. The results of the investigation certainly suggest further research. Such a classification as the Taxonomy should aid in clearly defining objectives and other behavioral outcomes such as the abilities involved in application, analysis, synthesis and evaluation should be studied. It is also suggested that the behaviors measured in this experiment, as well as others, be studied for longer periods of time than the period involved in this investigation.

CHAPTER V

SUMMARY

The purpose of this experiment was to study the differential retention of certain course outcomes in a beginning course in educational psychology. The course involved in the study was Education 62, Human Behavior and Development and the content area in which the different course outcomes were studied was tests and measurements.

Tests were constructed to measure four different behavioral objectives, namely: knowledge; translation; interpretation; and extrapolation. Items were devised to fit each of these tests by following the definitions as set forth and described in the "Taxonomy of Educational Objectives." The points of contact between the tests and the course were provided by a syllabus that had been jointly developed by the staff and used in Education 62 by all instructors.

A trial test was constructed and administered to a group of seventy-five students taking Education 62 the summer prior to the experiment. The results of the trial test were analyzed logically and statistically. Difficulty and discrimination of each item were studied and ineffective items were eliminated or changed. Items were also analyzed for evidence of ambiguities and ineffective foils. Further checks were made to assure consistency with behavioral definitions being measured. The reliability of the trial tests was computed and found to be satisfactory, although not high. Application of a departure from homogeneity test indicated heterogeneity of behaviors measured by most of the tests. The items were retained in their present classification, resulting in a refined

instrument with approximately twenty-four items in each test.

The refined tests were administered to a group of approximately 300 Education 62 students as a pretest, before studying course materials on tests and measurements, as a test at the end of the course and to a large proportion of this group approximately four months later.

The tests were again studied to determine their appropriateness for use in the retention experiment. Item difficulty and item discrimination were found to be satisfactory. Reliability of the tests was found to be low but acceptable. A further check was made to determine if each test was measuring a separate behavioral function. An F-test for departure from homogeneity indicated the desirability of combining the interpretation and extrapolation tests and further indicated that once these tests were combined separate functions were being measured by the resulting three tests.

Evidence of empirical validity was obtained by correlating the test scores with final course marks. The resulting correlations clustered about .57. Such a correlation gives some evidence of validity but also shows that there is considerable disparity in function of the tests and grades.

The degree to which the scholastic aptitude factor was represented in the various tests was determined by correlating the L-score of the American Council on Education Psychological Examination with each test. The resulting correlations hovered about .36 and were very nearly the same for all three tests.

The retention group was then studied. A t-test of significance between the means of the group of 172 students who took the retest and the group who did not was computed using the scores obtained at the time of

the second testing. The results indicated that the sample group was representative, showing a significant difference on only one of the tests, namely, translation.

The correlations of the test scores were reported between each administration of the tests. Results indicated a low positive relationship between the various administrations with slightly higher coefficients found for the interpretation-extrapolation test.

The differences between the mean scores of each administration were then tested using a t-test for correlated data. All t's were significant, showing that on all tests students learned a significant amount as demonstrated by the tests, retained a significant amount over what they knew at the time of the pretest and forgot a significant amount during the four-month retention period.

Graphic results of the retention data showed the largest amounts of material were learned in the knowledge area but also relatively larger amounts being forgotten. The relative amounts of gain and loss on the translation and interpretation-extrapolation tests were somewhat similar.

The percent of gain retained for each behavioral objective was determined. These percentages are as follows:

| <u>Test</u> | <u>Percent of gain retained</u> |
|----------------------------------|---------------------------------|
| Knowledge | 71.35 |
| Translation | 73.96 |
| Interpretation- Extrapolation | 63.57 |

When tests of significance were applied to the differences between these percentages it was found that knowledge and translation did not differ significantly, but that the interpretation-extrapolation test

scores differed from both knowledge and translation, significant beyond the one percent level of confidence.

It was concluded from this study that there is differential retention among the behavioral objectives measured, with the greatest degree of permanency being demonstrated in the area of interpretation and extrapolation.

SELECTED REFERENCES

1. Adams, G. S. and Torgersen, T. L. Measurement and Evaluation. New York: The Dryden Press, 1956.
2. American Council on Education Studies. A Design For General Education. Washington, D.C.: American Council on Education, June, 1944.
3. American Psychological Association. "Technical Recommendations for Psychological Tests and Diagnostic Techniques," Supplement to the Psychological Bulletin, Vol. 51, No. 2, Part 2, (March, 1954)
4. Dean, Kenneth L. Construction of Educational and Personnel Tests. New York: McGraw-Hill, 1953.
5. Bedell, Ralph C. "The Relationship Between the Ability to Recall and the Ability to Infer in Specific Learning Situations," Bulletin of the Northeast Missouri State Teachers College, LXXIV, No. 9, Kirksville: Northeast Missouri State Teachers College, 1934.
6. Bloom, Benjamin S. (ed.) Taxonomy of Educational Objectives. New York: Longmans, Green and Co., 1956.
7. Brownell, William A., Chairman, et al. "The Measurement of Understanding," The Forty-fifth Yearbook of the National Society for the Study of Education, Part I, Chicago: The University of Chicago Press, March, 1946.
8. Cronbach, Lee J. Essentials of Psychological Testing. New York: Harper and Brothers, 1949.
9. Burich, Alvin C. "Retention of Knowledge in a Course in General Psychology," Journal of Applied Psychology, Vol. 18 (April, 1934), pp. 209-19.
10. Fan, Chung-Teh. "Item Analysis Table." Princeton, New Jersey: Educational Testing Service, 1952.
11. Flanagan, John C. "The Use of Comprehensive Rationales in Test Development," Educational and Psychological Measurement, Vol. II, No. 1, 1951.
12. Prutchery, F. P. "Retention in High School Chemistry," Journal of Higher Education, Vol. 8 (1937), pp. 217-18.
13. Furst, Edward J. "Relationship Between Tests of Intelligence and Tests of Critical Thinking and Knowledge," Journal of Educational Research, Vol. 43, No. 8, (April, 1950), pp. 614-25.

14. Gerberich, J.R. Specimen Objective Test Items. New York: Longmans, Green and Co., 1956.
15. Greens, Edward B. "The Retention of Information Learned in College Courses," Journal of Educational Research, Vol. 24 (1931), pp. 262-72.
16. Greene, H. A., Jorgensen, A. V., and Gerberich, J. R. Measurement and Evaluation in the Secondary Schools. New York: Longmans, Green and Co., 1954.
17. Guilford, J. P. Fundamental Statistics in Psychology and Education. New York: McGraw-Hill Book Co., 1956.
18. Horrocks, John E. "The Relationship Between Knowledge of Human Development and the Ability to Use Such Knowledge," Journal of Applied Psychology, Vol. 30 (October, 1946), pp. 501-6.
19. Johnson, Palmer O. "Differential Functions of Examinations," Studies in College Examinations, Minneapolis: University of Minnesota, 1934, pp. 43-50.
20. Jordan, A. M. Measurement in Education. New York: McGraw-Hill, 1953.
21. Judd, C. H. Education as the Cultivation of Higher Mental Processes. New York: The MacMillan Co., 1936.
22. Lindquist, E. F. (ed.) Educational Measurement. Washington, D.C.: American Council on Education, 1951.
23. McWemar, Quinn. Psychological Statistics. New York: John Wiley and Sons, Inc., 1949.
24. McConnell, T. R. "A Study of the Extent of Measurement of Differential Objectives of Instruction," American Educational Research Association, (An Official Report), Washington, D.C.: American Educational Research Association, Feb., 1940, pp. 78-83.
25. Micheels, W. J. and Karnes, R. N. Measuring Educational Achievement. New York: McGraw-Hill, 1950.
26. Peters, C. C. "The Relation of Standardized Tests to Educational Objectives," Second Yearbook of the National Society for the Study of Educational Sociology. New York: Teachers College, Columbia University, 1929, pp. 148-59.
27. Pressey, Sidney L. and Robinson, Francis P. Psychology and the New Education. New York: Harper and Bros., 1944.
28. Remmers, H. H. and Gage, N. L. Educational Measurement and Evaluation. New York: Harper and Bros., 1955.

29. Ross, C. C. and Stanley, Julian C. Measurement in Today's Schools. New York: Prentice-Hall, Inc., 1954.
30. Smith, Eugene R., Tyler, Ralph W., and others. Appraising and Recording Student Progress. New York: Harper and Bros., 1942.
31. Stroud, James B. Psychology in Education. New York: Longmans, Green and Co., 1950.
32. Travers, Robert M. W. Educational Measurement. New York: The MacMillan Co., 1955.
33. Tyler, Ralph W. Basic Principles of Curriculum and Instruction. (Syllabus for Education 360), Chicago: University Press, 1950.
34. Tyler, Ralph W. Constructing Achievement Tests. Columbus, Ohio: Ohio State University, 1934.
35. Tyler, Ralph W. "Permanence of Learning," Journal of Higher Education, Vol. 4 (1933), pp. 203-4
36. Tyler, Ralph W. "What High School Pupils Forget," Educational Research Bulletin, Vol. 9 (1930), pp. 490-92.
37. Urdal, Lloyd B. "Interpretation of Factors by Means of a Taxonomy," Unpublished Doctoral thesis, Department of Education, University of Chicago, 1954.
38. Walker, Helen M. and Lev, Joseph. Statistical Inference. New York: Henry Holt and Co., 1953.
39. Ward, A. H. and Davis, R. A. "Individual Differences in Retention of General Subject Matter in the Case of Three Measurable Objectives," Journal of Experimental Education, Vol. 7 (1938), pp. 24-30.
40. Watson, Robert I. "An Experimental Study in the Permanence of Course Material in Introductory Psychology," Archives of Psychology, No. 225, 1938.
41. Wert, James E. "Twin Examination Assumptions," Journal of Higher Education, Vol. 8 (1937), pp. 136-40.
42. Wert, James E., Neidt, Charles O., and Ahman, J. S. Statistical Methods in Educational and Psychological Research. New York: Appleton-Century-Crofts, Inc., 1954.
43. Wrightstone, J. Wayne, Justman, Joseph, and Robbins, Irving. Evaluation in Modern Education. New York: American Book Co., 1956.

TRIAL TESTExamination

On

Tests and Measurements

OBJECTIVES OF EXAMINATION

This examination is designed to evaluate the student's abilities to deal with basic materials on tests and measurements in a beginning educational psychology course. The test situations are designed to measure the following course objectives:

1. The ability of the student to recognize or recall basic knowledge related to tests and testing.
2. The ability to translate this knowledge from one form into another to demonstrate an understanding of the knowledge.
3. The ability to interpret data relevant to the area of tests and measurements.
4. The ability to extrapolate from data, i.e. to go beyond the data and draw conclusions in line with the data presented.

DIRECTIONS

Each student will be provided with an answer sheet and a pencil. All responses are to be recorded on this answer sheet using the pencil provided. Make no marks on the booklet. You will find a set of instructions before each group of questions which you should read carefully. For each question in the list you are to choose the one best response. Work as rapidly as possible and answer all questions. Before beginning place your name and the name of your instructor in the space provided on your answer sheet. Also, please indicate on the back of your answer sheet those courses you intend to take next semester.

Directions for Multiple Choice Items: Choose the answer which you decide is correct and blacken the corresponding space on the answer sheet.

Part I - Knowledge

1. The major function of testing should be to
 - (1) make instruction less formal
 - (2) identify learning difficulties and successes of the pupils
 - (3) provide data for marking pupils
 - (4) determine the efficiency of individual teachers
2. Which of the following is most easily measured by a test?
 - (1) problem solving ability
 - (2) study skills
 - (3) factual information
 - (4) ability to comprehend
3. To obtain dependable evidence from any test one must
 - (1) convert the scores to percentiles or grades
 - (2) sample the student's performance in the specified area
 - (3) compare the performance to what others do on the same test
 - (4) keep a record to note degree of improvement over previous tests in this area
4. The best of the following criteria to determine if a test is good for your use is
 - (1) is it well recommended by the experts?
 - (2) does it fit the objectives of the course?
 - (3) does it sample all kinds of behavior?
 - (4) does it have alternate forms?
5. The standardized test as compared with the teacher-made test usually
 - (1) has a more specific purpose
 - (2) gives a better sample of school objectives
 - (3) is more valid because of objective scoring
 - (4) measures a wider scope of material
6. A serious weakness of formal tests is that they
 - (1) motivate students to learn the wrong thing
 - (2) are likely to obscure important school objectives
 - (3) have very little educative value
 - (4) encourage students to be dishonest
7. A generalization that might be made about most standardized tests is that they
 - (1) are difficult for the teacher to administer
 - (2) are relatively inappropriate for most things we do in school
 - (3) are misleading if treated as the sole evidence of merit
 - (4) usually require more time than can be justified as a part of any single course

8. A major use of standardized tests is to
- (1) help determine the student's grades
 - (2) compare your school with the neighboring cities
 - (3) motivate the student to work harder
 - (4) compare the performance of your students with norms
9. When a teacher wants to find out about a standardized test, what would be the best procedure?
- (1) write to the test company and ask for a write-up on the test
 - (2) contact the nearest university and ask if it is a good test
 - (3) consult Bureau of Mental Measurements Yearbook
 - (4) look through old college tests
10. Personality tests
- (1) depend largely upon the skill of the interpreter for their value
 - (2) have not yet proved their value in educational or vocational guidance
 - (3) are among the oldest of pupil appraisal tools
 - (4) usually possess a higher reliability than achievement tests
11. A survey test is a test that measures
- (1) specific strengths and weaknesses of a student in a given area
 - (2) general achievement of a group or an individual in a given subject or area
 - (3) what students know in all subjects or areas
 - (4) achievement only in the area of English
12. The test item you are now answering is an example of what type of item?
- (1) recognition
 - (2) recall
 - (3) subjective
 - (4) projective
13. The history of standardized testing goes back approximately how many years?
- (1) 30
 - (2) 50
 - (3) 70
 - (4) 90
14. A major objection to final examinations is that they
- (1) do not motivate students to study
 - (2) are a very poor sample of what the student knows about the subject
 - (3) are unfair to many students
 - (4) do not encourage self-evaluation
15. Which of the following devices would be of least value in making a judgment of a pupil's personality?
- (1) achievement tests
 - (2) projective techniques
 - (3) behavior diary records
 - (4) self-rating scales

16. Which of the following is an individual intelligence test?
(1) California Test of Mental Maturity
(2) Stanford Binet
(3) Otis Quick Score
(4) Primary Mental Abilities
17. The most appropriate kind of test to give a student who has a low mental age for his group would be
(1) written essay type test
(2) verbal test
(3) reading test
(4) performance test
18. The usual intelligence test best measures the capacity to learn
(1) art
(2) manipulative skills
(3) social skills
(4) verbal material
19. Most of our standardized intelligence tests assume that the student has had
(1) "normal" environmental background
(2) training in the same subjects in school
(3) no encounter with any situation on the tests
(4) opportunity to take some tests before
20. Which of the following best describes the accepted procedure in the use of the intelligence test results?
(1) give the I.Q. to parent and student
(2) never reveal the I.Q. to anybody
(3) reveal the interpretation of the I.Q. to the student and parent
(4) reveal the I.Q. to parents but not to students
21. In recent years the use of the I.Q. as a means of reporting performance has been replaced on many tests with the use of scores expressed as
(1) grade-equivalents
(2) percentile ranks
(3) grade placement
(4) educational ages
22. A test that places minor emphasis on the time limit is called a
(1) diagnostic test
(2) performance test
(3) survey test
(4) power test
23. A raw score is a score that
(1) has been converted to some standard scale for interpretation
(2) is an estimate of the student's performance on a test
(3) cannot be used in a distribution until it is changed
(4) shows the first quantitative results obtained in scoring a test

24. A student with an I.Q. of 84 would be classified as
- (1) average
 - (2) a moron
 - (3) low average
 - (4) feeble minded
25. Standardized achievement test results are most often reported in
- (1) grades - A, B, C, etc.
 - (2) raw scores
 - (3) grade placement scores
 - (4) quotient scores
26. When a test yields results consistently on each successive administration it is considered
- (1) reliable
 - (2) valid
 - (3) useable
 - (4) practical
27. The difference between the highest and lowest score in a distribution is called the
- (1) range
 - (2) spread
 - (3) deviation
 - (4) scatter
28. The point below and above which half of the test scores fall in a distribution is the
- (1) median
 - (2) mean
 - (3) mode
 - (4) center
29. The coefficient of correlation which shows the least amount of relationship is
- (1) 1.00
 - (2) .60
 - (3) .25
 - (4) -.35
30. The mean of a distribution is
- (1) the arithmetic average
 - (2) the mid-score
 - (3) another name for the median
 - (4) the same as the range

Part II - Translation

31. If scores on an intelligence test correlate .60 with success in college as measured by grades it means that
- (1) the abilities necessary to answer the intelligence test items are related to those necessary to get college grades
 - (2) 60 per cent of the material in the test is the same as that studied in college

- (3) there is practically no relationship between performance on this test and college success
 - (4) the test is right about 60 per cent of the time in predicting college grades
32. Sus was born July 9, 1948. What will her C.A. be on March 24, 1956?
- (1) 7-7
 - (2) 7-8
 - (3) 7-9
 - (4) 8-0
33. The accomplishment or achievement quotient is defined as the ratio of the educational age to the mental age. Which of the following would be the correct formula for this quotient?
- (1) $A.Q. = \frac{E.A.}{M.A.}$
 - (2) $A.Q. = \frac{M.A.}{E.A.}$
 - (3) $A.Q. = E.A. \times M.A.$
 - (4) none of the above
34. The I.Q. is the ratio of the mental age to the chronological age multiplied by 100. If you knew the chronological age and the I.Q. which of the following formulas would you use to find the mental age?
- (1) $M.A. = \frac{I.Q.}{C.A.} \times 100$
 - (2) $M.A. = \frac{C.A.}{I.Q.} \times 100$
 - (3) $M.A. = \frac{I.Q.}{C.A. \times 100}$
 - (4) $M.A. = \frac{I.Q. \times C.A.}{100}$
35. $I.Q. = \frac{M.A.}{C.A.}$ indicates the truth of which of the following statements?
- (1) The I.Q. will increase as the C.A. increases
 - (2) I.Q.'s at age 10 indicate greater variation in M.A.'s than they do at age 12
 - (3) if the M.A. changes the I.Q. will change
 - (4) if the ratio between M.A. and C.A. changes, the I.Q. will change
36. One objective of giving tests is to confirm or discourage a student's provisional tries; another way of saying this might be that a student
- (1) tries and if he fails he quits trying
 - (2) modifies his efforts according to the degree of success or failure on tests
 - (3) tries harder if he knows he is going to be tested and might fail
 - (4) gives up if he fails and continues to try harder if he succeeds
37. A major use of testing is for diagnosis. Which of the following represents an example of the foregoing statement?
- (1) a final examination
 - (2) a series of tests used to determine a student's grade
 - (3) an intelligence test
 - (4) an achievement battery early in the year

Listed below are several test situations (Nos. 38-47) which might appear on different kinds of standardized tests. On your answer sheet, if the item would most likely appear on
 an intelligence test, mark (1)
 a special ability test, mark (2)
 an achievement test, mark (3)
 an interest test, mark (4)
 a personality test, mark (5)

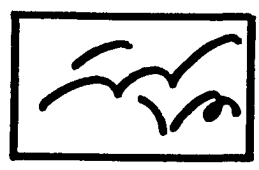
- 38. Repeat backwards "4-7-6-3-2"
- 39. A preface is found in what part of the book or chapter?
 (A) beginning
 (B) middle
 (C) end
- 40. 6, 4, 7, 5, 8, 6, 9, — What number should come next?
 (A) 7 (B) 10 (C) 8 (D) 6 (E) 11



- 42. Tell the one you like least and the one you like most:
 (A) Develop new varieties of flowers
 (B) Conduct advertising campaign for florists
 (C) Take telephone orders in a florist shop
- 43. Choose one of the following:
 (A) I wish I didn't have so many aches and pains
 (B) I wish I wouldn't keep changing my mind
- 44. The child is given "colored mud" and is allowed to make designs or pictures, or just to enjoy manipulating it.
- 45. Which word does not belong with the others?
 (A) apparatus
 (B) foundation
 (C) equipment
 (D) device
 (E) appliance
- 46. Find the area of a triangle having a base of 20 inches and an altitude of 12 inches.
- 47. Which of the following pictures is more appealing?



(A)



(B)

Examine the graph and answer the questions that follow. (Nos. 48-51).

Changes in Mental Ability With Age on the
Wechsler Bellevue Intelligence Test



48. The average performance begins to drop off in the
- (1) early teens
 - (2) late teens
 - (3) early twenties
 - (4) late twenties
49. The performance on this test at the age of 65 is approximately equal to the performance of persons at the age of
- (1) 10
 - (2) 12
 - (3) 15
 - (4) 17
50. Growth in ability according to this test is most rapid during which of the following periods?
- (1) 13 to 15
 - (2) 15 to 17
 - (3) 17 to 19
 - (4) 19 to 21
51. We could conclude from these data that
- (1) some teenagers are smarter than older people
 - (2) older people are just as smart but don't show it on the test
 - (3) the Wechsler is not a good test for adults
 - (4) some teenagers make higher Wechsler scores than do older people
-
52. When we say that standardized tests enable a teacher to evaluate more objectively the abilities of a student, we mean that
- (1) the teacher can verify her judgment by using an unbiased tool
 - (2) the student then becomes the object of the evaluation
 - (3) the abilities of each student can be measured only by an expert and the expert makes up the test
 - (4) objective qualities appear on the test that the teacher didn't know about

53. An appropriate test is said to have curricular validity. Which of the following testing situations would be most likely to have this characteristic?
- (1) a personality test in a physics course
 - (2) a composition test in a literature course
 - (3) repairing a broken tool in a shop mechanics course
 - (4) a test of facts and knowledge in a home economics course
54. Norms serve as basis for interpreting scores of the individual or class means that, they
- (1) determine if class behavior is normal
 - (2) indicate what the class or student should do
 - (3) serve as a basis for passing or failing students
 - (4) indicate what the average pupils do

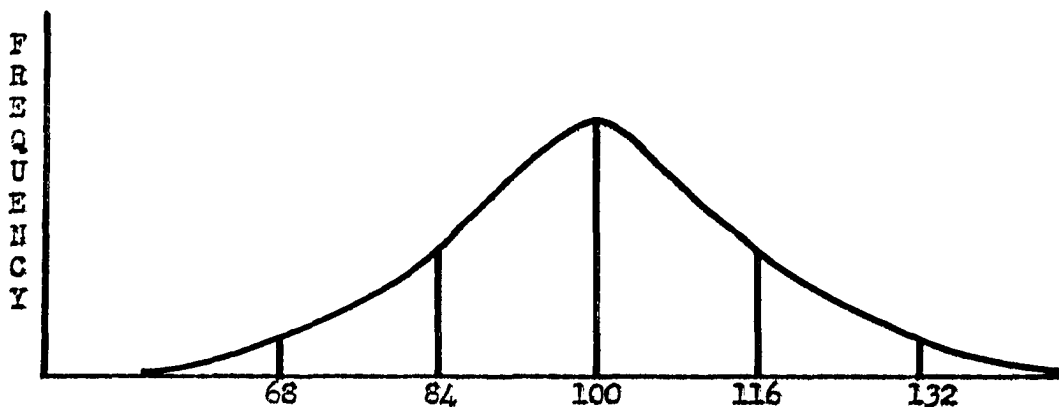
Part III - Interpretation and Extrapolation

Examine the data in the following table and answer questions Nos. (55-58). Answers are to be determined on basis of data alone.

| <u>Mental Age Range by School Grade</u> | |
|---|--|
| <u>Grade</u> | <u>M.A. Range (2nd to 98th percentile)</u> |
| 11 | 8.4 |
| 9 | 8.4 |
| 7 | 7.2 |
| 5 | 5.6 |
| 3 | 4.8 |
| 1 | 3.6 |

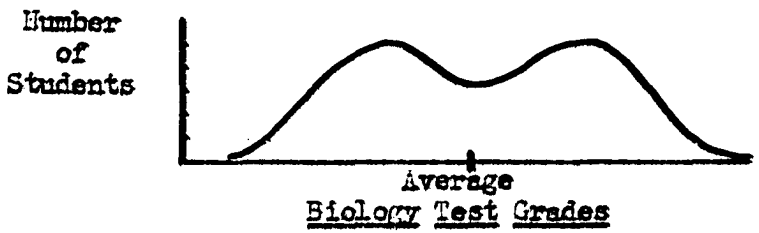
55. A teacher would have her greatest problem of individual differences at what grade level?
- (1) third
 - (2) fifth
 - (3) seventh
 - (4) ninth
56. What would be your best estimate of the M.A. range in grade 6? (2nd to 98th percentile)
- (1) 6.0
 - (2) 6.4
 - (3) 6.8
 - (4) 7.2
57. A high school teacher will find differences between the extremes of the mental ages of approximately
- (1) 4 to 6 years
 - (2) 6 to 8 years
 - (3) 8 to 10 years
 - (4) 10 to 12 years
58. What would be your best estimate to the nearest year of the M. A. range in kindergarten? (2nd to 98th percentile)
- (1) 2 years
 - (2) 3 years
 - (3) 4 years
 - (4) 5 years

Study the curve given and answer the questions following. (Nos. 59-64).



59. According to this curve which is the most common score among the following?
- (1) 70
 - (2) 130
 - (3) 86
 - (4) 112
60. We would expect that 95-100% of the I.Q.'s would fall below
- (1) 84
 - (2) 100
 - (3) 116
 - (4) 132
61. The number of people getting I.Q.'s of 140 would be equal to the number getting I.Q.'s of
- (1) 60
 - (2) 68
 - (3) 100
 - (4) 132
62. The top I.Q. according to this curve would be
- (1) 132
 - (2) 140
 - (3) 150
 - (4) impossible to tell
63. The greatest number of people would fall in which of the following I.Q. ranges
- (1) 68-84
 - (2) 116-132
 - (3) 132 and above
 - (4) 84-92
64. You would expect approximately what per cent of the people to have I.Q.'s of less than 50?
- (1) .5% or less
 - (2) 2%
 - (3) 5%
 - (4) 10%

Mr. Smith gave a biology test in his class, a typical sophomore group. He drew a curve showing the distribution of the test scores. Refer to this curve and answer the questions following. (Nos. 65-67).



- 65. We would expect to find that the test scores indicated
 - (1) about the same number of high and low grades
 - (2) more high than low grades
 - (3) more low than high grades
 - (4) most of the grades around the average

- 66. When Mr. Smith assigned grades, he would likely have
 - (1) more A's than F's
 - (2) more D's than C's
 - (3) more C's than B's
 - (4) more C's than D's plus B's

- 67. The best guess we could make about Mr. Smith's students with regard to time the students studied for the test is
 - (1) they all studied very hard for the test
 - (2) some studied and some didn't but most of them did
 - (3) many studied and many didn't study
 - (4) the average student studied pretty hard

Data are given below on five pupils enrolled in a class of 30 ninth graders. The test data are based on performance at the end of the first semester. You are to read over the summary and then show to which pupil each statement best fits by marking the pupil's number of the answer sheet. (Nos. 68-72).

| Pupil | I.Q. | Calif. Ach. Test Performance | | | Teacher's estimate of Ach. Rank in Class |
|-------|------|------------------------------|-------|-------|--|
| | | Arith. | Read. | Lang. | |
| 1 | 88 | 9.1 | 8.0 | 8.3 | 20 |
| 2 | 99 | 9.7 | 9.6 | 9.5 | 14 |
| 3 | 132 | 9.5 | 9.8 | 10.2 | 12 |
| 4 | 138 | 11.8 | 12.3 | 12.0 | 3 |
| 5 | 101 | 10.0 | 10.1 | 10.9 | 4 |

- 68. The pupil who should be doing considerably better in his school achievement.
- 69. The accuracy of the I.Q. seems most doubtful in which case?
- 70. A bright student making good use of his ability.
- 71. Teacher regards aptitude too highly according to test results.

72. Teacher's rank most consistent with test scores.

The following test scores are available on Tom, a senior in high school. Look them over carefully and answer the questions that follow. (Nos. 73-75).

Terman-McNemar Test of Mental Ability-Age 15-0; I.Q. 143

Kuder Preference Record

| <u>Significantly high</u> | <u>Definitely low</u> |
|---------------------------|-----------------------|
| Computational | Social Science |
| Scientific | Clerical |
| Literary | |

Heston Personal Adjustment Inventory - Senior Norms

| | |
|---------------------|----|
| Analytical Thinking | 96 |
| Home Satisfaction | 70 |
| Emotional Stability | 60 |
| Sociability | 8 |
| Confidence | 12 |
| Personal Relations | 6 |

Refer only to the above test results and answer the following questions:

73. Tom's ability is best described as

- (1) above average
- (2) superior
- (3) very superior
- (4) high genius

74. Tom's scores indicate that he would be best suited for

- (1) research
- (2) medicine
- (3) teaching
- (4) selling

75. Tom's test score patterns indicate a need to

- (1) widen his scope of interest
- (2) make better use of his ability
- (3) set a definite goal
- (4) improve his sociability

Mr. Tuttle found that his norms did not go high enough to interpret the test score of one of his students. The last four norms are shown below but Sally got a score of 130.

| <u>Score</u> | <u>Age Equivalent</u> |
|--------------|-----------------------|
| 120 | 12-6 |
| 115 | 12-2 |
| 110 | 12-0 |
| 105 | 11-6 |

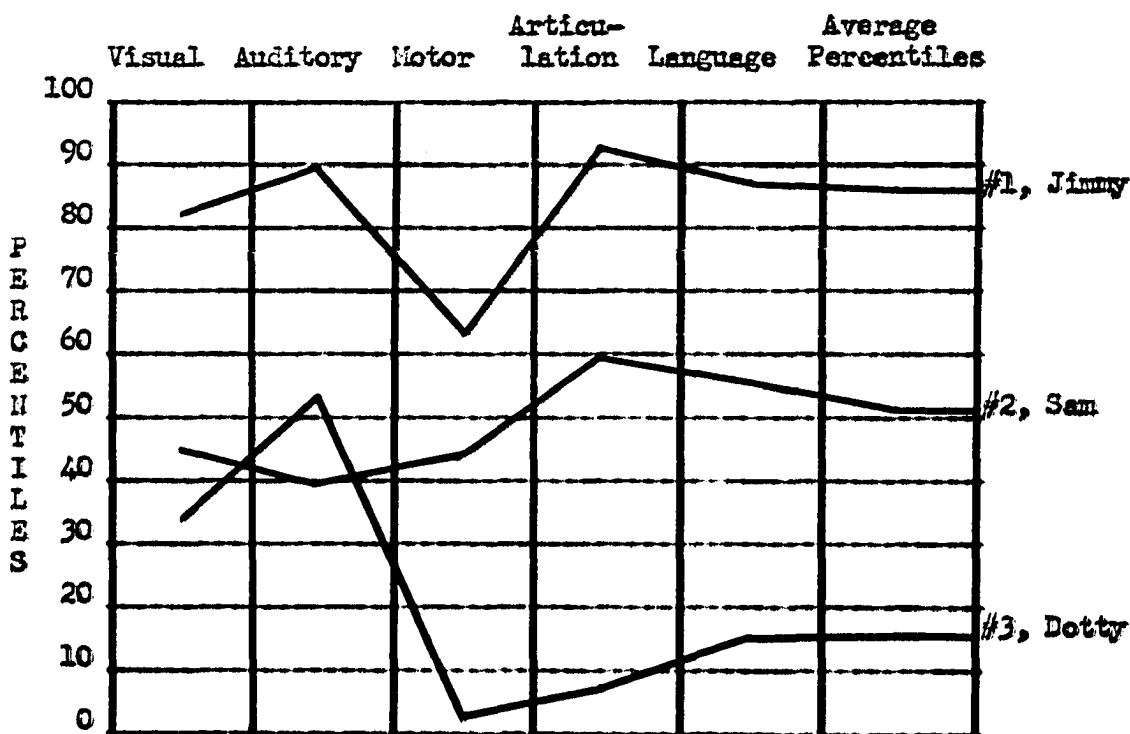
76. What would the best estimate of Sally's grade equivalent be?

- (1) 13-0
- (2) 13-1
- (3) 13-2
- (4) over 12-6

Examine the data and profiles of these three entering first graders, and answer questions that follow. (Nos. 77-83).

1. Jimmy, age 6-0, M.A. 7-4
2. Sam, age 6-1, M.A. 7-5
3. Dotty, age 6-3, M.A. 5-3

Reading Readiness Abilities



On your answer sheet mark

- (1) if the statement is most true of #1, Jimmy.
- (2) if the statement is most true of #2, Sam.
- (3) if the statement is most true of #3, Dotty.
- (4) if the statement does not validly apply to any of the three students.

77. Is at the kindergarten level mentally.
78. Requisites best developed for participation in beginning reading experiences.
79. Teacher will most likely provide at once activities and exercises to develop latent motor ability.
80. Readiness scores most out of line with mental ability.
81. Should have immediate examination by ear specialist.
82. Poorly developed abilities for first grade activities.
83. Apparently very well adjusted socially.

Susan's record shows that she has taken two achievement batteries. The first one she took at the beginning of the eighth grade and the other in the middle of the tenth grade. Examine the results of these two batteries and answer the questions following. (Nos. 84-89).

8th grade, Progressive Achievement Tests, Intermediate Battery

| | <u>Grade Placement</u> |
|-------------------------|------------------------|
| Reading Vocabulary | 9.3 |
| Reading Comprehension | 7.6 |
| Total Reading | 8.5 |
| Arithmetic Reasoning | 7.0 |
| Arithmetic Fundamentals | 9.0 |
| Total Arithmetic | 8.2 |
| Language | 8.6 |
| Total for Battery | 8.7 |

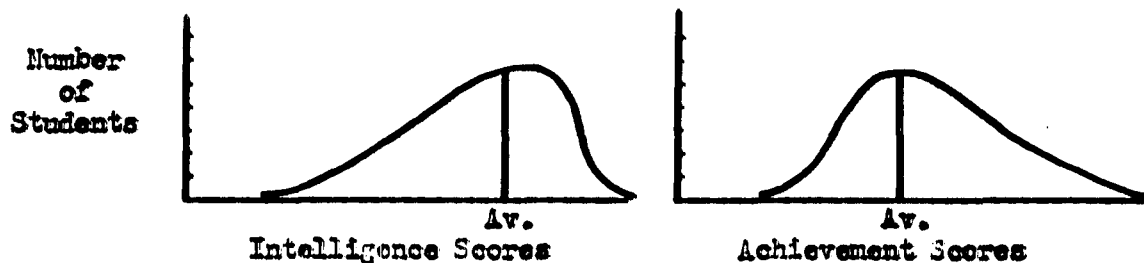
10th grade, Iowa Tests of Ed. Development

| | <u>Percentile Scores</u> |
|--|--------------------------|
| Understanding of basic social concepts | 58 |
| Background of natural science | 54 |
| Correctness of writing | 73 |
| Quantitative Thinking | 71 |
| Ability to interpret reading materials in the social studies | 51 |
| Ability to interpret reading materials in the natural sciences | 46 |
| Ability to interpret literary materials | 62 |
| General Vocabulary | 74 |
| Uses of sources of information | 78 |

84. There is most disagreement on the two test batteries in the area of
- (1) ability to think quantitatively
 - (2) knowledge of fundamentals
 - (3) vocabulary
 - (4) ability to read and interpret reading
85. If we were to administer another test such as the California Test of English Usage at the beginning of grade 11, what approximate grade placement would you predict from the previous scores?
- (1) 10.0
 - (2) 11.0
 - (3) 12.0
 - (4) 13.0
86. Judging from the test results only we might infer that Susan came from what type of home environment?
- (1) unstimulating and deprived atmosphere
 - (2) a home that emphasized deep thinking and problem solving
 - (3) a stimulating environment with much opportunity to read
 - (4) a good home physically, but with parents that didn't care about schooling

87. Judging on the basis of the total test results which general ability seems to be the strongest? Ability to
- (1) deal with specifics
 - (2) interpret and solve problems
 - (3) deal with quantitative material
 - (4) draw conclusions
88. On the basis of these test results how well would you expect Susan to do in a course in high school physics?
- (1) below average
 - (2) average
 - (3) above average
 - (4) very well
89. How does the over-all test performance of the Iowa Test compare with the Progressive Test?
- (1) much lower
 - (2) about the same
 - (3) much higher
 - (4) can't be compared

Study the curves given and answer the questions following. (Nos. 90-93).



Curve # 1 shows the distribution of I.Q.'s in a class
 Curve # 2 shows the distribution of achievement test scores in the same class

90. The class, as a group, appears to be
- (1) underachievers
 - (2) overachievers
 - (3) achieving in line with their ability
 - (4) it is impossible to determine their rate of achievement
91. A student who had an average I.Q. in this class would probably
- (1) have average achievement
 - (2) have slightly above average achievement
 - (3) have slightly below average achievement
 - (4) have very low achievement
92. The range of intelligence test scores as compared to the range of achievement test scores
- (1) tends to be greater
 - (2) tends to be less
 - (3) tends to be about the same
 - (4) can't tell from the data given

93. A best guess with regard to the students of this class would be that they
- (1) are having too many assignments
 - (2) are not motivated
 - (3) are doing about what we would expect
 - (4) will do better on the next test

| | | <u>Reading</u> | | | | |
|--|---------------------------------|-------------------------------|-------------------|--------------------|---------------------|---------------------------------|
| | | Two yrs. Behind or more | One yr. Behind | Normal | One yr. Advanced | Two yrs. Advanced or more |
| A r i t h m e t i c | Two yrs. Advanced or more | | | x x | x | x x |
| | One yr. Advanced | | x | x x x | x x x | x x x |
| | Normal | | x x x | x x x x x x x x | x x x | |
| | One yr. Behind | x x x | x x x | x x x | x | |
| | Two yrs. Behind or more | x x | x | x x | | |
| | | Average and Retarded Readers | | | Superior Readers | |

Making your judgment on the basis of the information given in the graph, classify each of the following by marking Nos. 94-101.

- (1) if the item is definitely true
 - (2) if the item is probably true
 - (3) if the information given is insufficient to make a judgment regarding the truth or falsity of the item
 - (4) if the item is probably false
 - (5) if the item is definitely false
94. In general, children who are good readers are likely to be good in arithmetic.
95. If we found a student that was advanced in reading we could conclude for most purposes that he would be good in all subjects.
96. Of the group of superior readers there are four who are retarded in arithmetic.
97. There is a higher relationship between reading and arithmetic than there is between reading and other subjects.
98. The best prediction we could make about students two years advanced in reading is that they would be two years advanced in arithmetic.
99. The chart offers evidence for ability grouping.
100. The cluster of eight students in the center of the graph implies that ability is the same from one subject to the next.
101. There are 22 students considered on this graph.

Questions 102-111 refer to the following data. Examine these data and answer the questions that follow. Directions are given following the data.

Test data taken from the record of Bill who is at present a junior in high school and is 16 years and 3 months old.

Otis Quick Score Test of Mental Ability - 3rd grade I.Q. 104

California Test of Mental Maturity - 7th grade

Verbal I.Q. 116
Non-Verbal I.Q. 104
Total I.Q. 112

Chicago Tests of Primary Mental Abilities - 9th grade

| | Percentile |
|----------------|------------|
| Verbal Meaning | 71 |
| Reasoning | 73 |
| Space | 53 |
| Number | 45 |
| Word Fluency | 62 |
| Total | 65 |

Kuder Preference Record - 11th grade.

| | | Percentile scores listed. |
|---------------|----|---------------------------|
| Mechanical | 32 | Social Service 88 |
| Persuasive | 60 | Scientific 72 |
| Musical | 23 | Literary 90 |
| Computational | 36 | Clerical 62 |
| Artistic | 22 | |

California Achievement Test, Complete Battery, beginning of grade 9, grade placement listed.

| | |
|-------------|------|
| Reading | 11.3 |
| Language | 11.1 |
| Mathematics | 9.8 |
| Total | 10.8 |

The principal of his school looked over these data and drew several conclusions. Some of these conclusions are listed below together with comments other principals or teachers might make. For each of these conclusions check the comment which might be most appropriately made about this conclusion. Base your conclusions on the test data only.

102. "Bill's intelligence is above average."

- (1) This conclusion is true
- (2) This is probably true but we can't be sure because the sample of his intelligence is inadequate to draw a final conclusion
- (3) This is probably true because he is low in mathematics
- (4) This would be true only if we assure that some of his scores are incorrect

103. "Bill is interested in the subjects he is best in."

- (1) This conclusion is true
- (2) This is probably true but we can't be sure because of scanty evidence
- (3) This is probably untrue because of conflicting evidence on his test scores
- (4) This is false as the test results indicate

104. "I'm sure that something went wrong when Bill took the math test."
(1) This conclusion is true
(2) This conclusion is probably true but one cannot be sure because his computational interest is low
(3) This conclusion is probably untrue because scores seem to indicate Bill has less ability in this area.
(4) This statement is false because Bill did better than should be expected in mathematics
105. "Bill's over-all achievement is about what it should be."
(1) This statement is true
(2) This statement is probably true but we can't be sure because of the low scores on the Kuder
(3) This statement is probably not true because of the low score on math
(4) Statement is not true as the achievement scores are out of line with the ability scores
106. "Bill's strength seems to lie in the verbal area."
(1) This conclusion is true
(2) This conclusion probably is true providing we disregard the Chicago tests
(3) This conclusion is probably false since the Otis score is so low
(4) This conclusion is not true
107. "Bill could be advanced a grade if these scores are correct."
(1) Statement is true
(2) Statement is probably true but we can't be sure since the scores offer conflicting evidence
(3) Statement is probably untrue because the scores are mostly in line with his present grade placement
(4) Statement is false as there is no indication he could do the work in the next grade.
108. "We might expect Bill's I.Q. to increase from the third grade."
(1) Statement is true
(2) Statement is probably true because of the other types of ability scores given
(3) Statement is probably untrue because the Chicago test scores bear out the third grade I.Q.
(4) Statement is not true as there is not enough evidence to indicate a significant change
109. "Bill will do well in college."
(1) Conclusion is true
(2) Conclusion is probably true since the verbal interests seem to be high
(3) Conclusion is probably not true because of the other scores given on Bill
(4) Conclusion is not warranted on basis of these scores

In addition answer the following question pertaining to the test data regarding Bill on page 17, Nos. 110-111.

110. The percentile score of 88 in Social Service on the Kuder indicates
- (1) that Bill got 88% of the answers correct
 - (2) that he has more ability in Social Service than 88% of his norm group
 - (3) that only 12% of the norm group showed more interest in social studies than he did
 - (4) that 88 out of 100 will do better than he did on the test
111. The grade placement of 9.8 in mathematics means that
- (1) Bill's achievement is equivalent to the average ninth grader who has been in school 8 months
 - (2) Bill's achievement is slightly below average for the tenth grade in math
 - (3) Bill's grade placement in math is correct considering his age
 - (4) Both (1) and (2) are correct

The five students for whom the data are given below are in kindergarten. These test data are based on test performance at the beginning of the second semester. After examining the data indicate which pupil best fits each of the following statements by marking the number of the student on the answer sheet. (Nos. 112-118).

| <u>Student</u> | <u>C.A.</u> | <u>I.I.A. on Stanford Binet</u> | <u>Percentile Rank On Readiness Test</u> |
|----------------|-------------|---------------------------------|--|
| 1 | 5-10 | 7-4 | 72 |
| 2 | 6-4 | 5-4 | 22 |
| 3 | 5-10 | 5-5 | 64 |
| 4 | 5-8 | 5-6 | 45 |
| 5 | 5-6 | 6-10 | 38 |

Which student

112. Is most ready at present for first grade work?
113. Is apparently in need of stimulating experiences but has fairly high aptitude?
114. Will probably not be able to keep up with the average in the first grade?
115. Apparently comes from a very stimulating environment?
116. Is most characteristic of the average for this group?
117. Can you predict will have the lowest ability three years from this time?
118. Is least ready for first grade work?

APPENDIX B

Item Difficulty and Item Discrimination
For the Trial Test Scores

| Test | Item Number | Percent Correct Responses | Correlation with Test Score |
|-------------|-------------|---------------------------|-----------------------------|
| Knowledge | 1 | -.26 | 98.6 |
| | 2 | .84 | 64.0 |
| | 3 | .00 | 4.1 |
| | 4 | .25 | 90.6 |
| | 5 | .05 | 48.0 |
| | 6 | .38 | 73.3 |
| | 7 | .66 | 86.6 |
| | 8 | .15 | 94.7 |
| | 9 | .51 | 44.0 |
| | 10 | .72 | 74.7 |
| | 11 | .51 | 46.7 |
| | 12 | .52 | 56.0 |
| | 13 | .21 | 61.3 |
| | 14 | .24 | 68.0 |
| | 15 | .66 | 77.3 |
| | 16 | .31 | 51.3 |
| | 17 | .21 | 78.7 |
| | 18 | .55 | 79.7 |
| | 19 | .21 | 57.3 |
| | 20 | .49 | 53.3 |
| | 21 | .05 | 68.0 |
| | 22 | .36 | 47.9 |
| | 23 | .51 | 72.0 |
| | 24 | .21 | 85.3 |
| | 25 | .25 | 81.3 |
| | 26 | -.10 | 82.7 |
| | 27 | .07 | 73.3 |
| | 28 | .62 | 81.3 |
| | 29 | .59 | 38.8 |
| | 30 | .60 | 46.7 |
| Translation | 31 | .55 | 44.6 |
| | 32 | .40 | 18.9 |
| | 33 | .36 | 54.1 |
| | 34 | .40 | 50.0 |
| | 35 | .28 | 70.7 |
| | 36 | .44 | 77.3 |
| | 37 | .32 | 72.0 |
| | 38 | .73 | 51.3 |
| | 39 | .64 | 43.4 |

| Test | Item Number | Percent Correct Responses | Correlation with Test Score |
|---------------------|-------------|---------------------------|-----------------------------|
| Translation (Cont.) | 40 | .72 | 70.3 |
| | 41 | .17 | 26.7 |
| | 42 | .45 | 69.3 |
| | 43 | .59 | 82.7 |
| | 44 | .32 | 35.1 |
| | 45 | .32 | 57.5 |
| | 46 | .45 | 60.3 |
| | 47 | .43 | 15.1 |
| | 48 | .21 | 42.7 |
| | 49 | .15 | 62.7 |
| | 50 | .00 | 58.7 |
| | 52 | .21 | 62.7 |
| | 53 | .56 | 52.7 |
| | 54 | -.06 | 76.0 |
| | 59 | .51 | 77.0 |
| | 60 | .33 | 67.1 |
| | 65 | .68 | 60.5 |
| | 73 | .49 | 45.3 |
| | 77 | .35 | 82.7 |
| | 96 | .68 | 75.7 |
| 101 | .42 | 60.0 | |
| 110 | .33 | 36.1 | |
| 111 | .30 | 21.9 | |
| Interpretation | 51 | .11 | 64.5 |
| | 62 | .59 | 80.0 |
| | 63 | .30 | 85.3 |
| | 68 | .62 | 77.6 |
| | 69 | .56 | 32.9 |
| | 70 | .48 | 76.0 |
| | 71 | .21 | 40.0 |
| | 72 | .31 | 48.6 |
| | 78 | .56 | 81.1 |
| | 80 | .60 | 60.8 |
| | 81 | .48 | 22.7 |
| | 82 | .35 | 86.7 |
| | 83 | .56 | 33.3 |
| | 84 | .48 | 20.0 |
| | 87 | .21 | 35.1 |
| | 89 | .15 | 44.6 |
| | 90 | .63 | 33.3 |
| | 91 | .40 | 31.1 |
| | 92 | .30 | 20.3 |
| | 97 | .69 | 42.7 |
| 102 | -.18 | 27.0 | |
| 103 | .26 | 56.8 | |
| 106 | .38 | 84.0 | |
| 108 | .16 | 37.0 | |
| 112 | .68 | 84.9 | |
| 118 | .53 | 95.7 | |

| Test | Item Number | Percent Correct Responses | Correlation with Test Score |
|---------------|-------------|---------------------------|-----------------------------|
| Extrapolation | 55 | .36 | 52.1 |
| | 56 | .18 | 79.7 |
| | 57 | .33 | 27.0 |
| | 58 | .51 | 35.6 |
| | 61 | .58 | 87.8 |
| | 64 | .31 | 52.6 |
| | 66 | .63 | 38.7 |
| | 67 | .73 | 47.4 |
| | 74 | .62 | 90.7 |
| | 75 | .43 | 83.8 |
| | 76 | .26 | 41.3 |
| | 79 | .33 | 31.1 |
| | 85 | .33 | 30.1 |
| | 86 | .45 | 42.7 |
| | 88 | .36 | 37.3 |
| | 93 | .31 | 49.3 |
| | 94 | .16 | 34.7 |
| | 95 | .59 | 32.4 |
| | 98 | -.18 | 18.4 |
| | 99 | .21 | 21.6 |
| | 100 | .44 | 22.7 |
| 104 | .59 | 72.6 | |
| 105 | -.21 | 46.7 | |
| 107 | -.12 | 34.7 | |
| 109 | .07 | 33.3 | |
| 113 | .65 | 73.6 | |
| 114 | .48 | 81.1 | |
| 115 | .50 | 27.4 | |
| 116 | .70 | 62.4 | |
| 117 | .48 | 88.7 | |

REQUIRED TEST

EXAMINATION ON TESTS AND MEASUREMENTS

Purpose of Examination

This four-part examination is designed to evaluate the following abilities of the student in the area of tests and measurements.

The ability to:

1. Recognize or recall basic knowledge related to tests and measurements.
2. Translate such knowledge from one form into another to demonstrate an understanding of it.
3. Interpret data relevant to the area of tests and measurements.
4. Extrapolate from data, i.e. to go beyond the data and draw conclusions from it.

Directions

Each student will be provided with an answer sheet and a pencil. All responses are to be recorded on this answer sheet using the pencil provided. Make no marks on the question booklet. Fill in the information requested on your answer sheet and also read directions given for marking it. For each question in the list you are to choose the one best response. Work as rapidly as possible and answer all questions.

Part I - Knowledge

1. Which of the following is most easily measured by a test?
 - (1) problem-solving ability
 - (2) study skills
 - (3) factual information
 - (4) ability to comprehend
2. In selecting a standardized test for use in any course, the first consideration should be if the test is:
 - (1) long enough to sample all kinds of behavior in the course
 - (2) well recommended by the authorities in this subject matter area
 - (3) practical for use in the course, i.e. easy to give, score, interpret, etc.
 - (4) fitted to the particular objectives of the course
3. A serious weakness of formal tests is that they:
 - (1) motivate students to learn the wrong thing
 - (2) are likely to obscure important school objectives
 - (3) have very little educative value
 - (4) over-emphasize the student's ability to solve problems
4. A generalization that might be made about most standardized tests is that they:
 - (1) are difficult for the teacher to administer
 - (2) are relatively inappropriate for most things we do in school
 - (3) are misleading if treated as the sole evidence of merit
 - (4) usually require more time than can be justified as a part of any single course


5. When a teacher wants to find out about a standardized test, what is the best procedure?
- (1) write to the test company and ask for a write-up on the test
 - (2) contact the nearest university and ask if it is a good test
 - (3) consult Buros Mental Measurements Yearbook
 - (4) look through college texts containing test information
6. Personality tests
- (1) depend largely upon the skill of the interpreter for their value
 - (2) usually possess a higher reliability than achievement tests
 - (3) are among the oldest of pupil appraisal tools
 - (4) have not yet proved their value in educational or vocational guidance
7. A survey test is a test that measures:
- (1) specific strengths and weaknesses of a student in a given area
 - (2) general achievement of a group or an individual in a given subject or area
 - (3) what students know in all subjects or areas
 - (4) a pupil's performance of a complex task broken down into several parts
8. A major objection to final examinations is that they:
- (1) do not measure what is taught in the course
 - (2) are a very poor sample of what the student knows about the subject
 - (3) are unfair to many students
 - (4) do not encourage self-evaluation
9. The test item you are now answering is an example of what type of item?
- (1) recognition
 - (2) recall
 - (3) subjective
 - (4) projective
10. Which of the following devices would be of least value in making a judgment of a pupil's personality?
- (1) achievement tests
 - (2) projective techniques
 - (3) behavior diary records
 - (4) self-rating scales
11. Which of the following is an individual intelligence test?
- (1) California Test of Mental Maturity
 - (2) Stanford Binet
 - (3) Ohio State Psychological Test
 - (4) Primary Mental Abilities
12. The usual intelligence test best measures the capacity to learn which of the following skills?
- (1) quantitative
 - (2) manipulative
 - (3) social
 - (4) verbal
13. Most of our standardized intelligence tests assume that the student has had:
- (1) "normal" environmental background
 - (2) training in the same subjects in school as other students
 - (3) no previous encounter with any situation on the tests
 - (4) average social intelligence
14. A test that places minor emphasis on the time limit is called a:
- (1) diagnostic test
 - (2) performance test
 - (3) survey test
 - (4) power test

15. Which of the following would be of most value in determining the typical behavior of a student?
- (1) observation
 - (2) projective testing
 - (3) individual intelligence testing
 - (4) school achievement records
16. Which of the following best describes the accepted procedure in the use of the intelligence test results?
- (1) give the I.Q. to parents and student if they request it and seem serious about the matter
 - (2) never reveal the I.Q. to anybody
 - (3) reveal the I.Q. to parents but not the student
 - (4) reveal the interpretation of the I.Q. to the parents or student
17. A raw score is a score that:
- (1) shows a certain percent of achievement on a given test
 - (2) is an estimate of the student's performance on a test
 - (3) cannot be used in a distribution until it is changed
 - (4) shows the first quantitative results obtained in scoring a test
18. Standardized achievement test results are most often reported in:
- (1) standard scores
 - (2) raw scores
 - (3) grade placement scores
 - (4) quotient scores
19. The score which is attained by the greatest number in any group of scores is called the:
- (1) mean
 - (2) median
 - (3) mode
 - (4) midscore
20. The point below and above which half of the test scores fall in a distribution is the:
- (1) mean
 - (2) median
 - (3) mode
 - (4) range
21. Which of the following correlation coefficients shows the least amount of relationship?
- (1) 1.00
 - (2) .60
 - (3) .25
 - (4) -.35
22. The mean of a distribution is:
- (1) the mid-score
 - (2) the arithmetic average
 - (3) another name for the median
 - (4) the same as the range
23. The average score obtained on a test by pupils of a given grade placement is a grade:
- (1) quotient
 - (2) norm
 - (3) score
 - (4) rank
24. If the primary function of the test is to help predict and plan subsequent school work for the student it would be called what kind of a test?
- (1) prognostic
 - (2) diagnostic
 - (3) power
 - (4) therapeutic

Part II - Translation

Listed below are several test situations which might appear on different kinds of standardized tests. (Numbers 25 to 33) On your sheet, if the item would most likely appear on:

an intelligence test, mark (1) an interest test, mark (4)
 a special ability test, mark (2) a personality test, mark (5)
 an achievement test, mark (3)

25. Repeat backwards, "4-7-6-3-2."
26. A preface is found in what part of the book or chapter?
 (A) beginning
 (B) middle
 (C) end
27. 6,4,7,5,8,6,9,— What number should come next?
 (A) 7 (B) 10 (C) 8 (D) 6 (E) 11
28. Tell the one you like least and the one you like most:
 (A) Develop new varieties of flowers
 (B) Conduct advertising campaign for florists
 (C) Take telephone orders in a florist shop
29. Choose one of the following:
 (A) I wish I didn't have so many aches and pains
 (B) I wish I wouldn't keep changing my mind
30. The child is given "colored mud" and is allowed to make designs or pictures, or just to enjoy manipulating it.
31. Which word does not belong with the others?
 (A) apparatus
 (B) foundation
 (C) equipment
 (D) device
 (E) appliance
32. Find the area of a triangle having a base of 20 inches and an altitude of 12 inches.
33. Which of the following designs is more appealing? 
-
34. If a beginning ninth grader achieves a grade placement of 9.8 on a mathematics achievement test, it means that:
 (1) his achievement is equivalent to the average ninth grader who has been in school eight months
 (2) his achievement is slightly below average for the tenth grade in math
 (3) he should be achieving at the 9.8 level in math
 (4) both (1) and (2) are correct

35. If scores on an intelligence test correlate .60 with success in college as measured by grades it means that:
- (1) the abilities necessary to answer the intelligence test items are related to those necessary to get college grades
 - (2) 60 percent of the material in the test is the same as that studied in college
 - (3) there is practically no relationship between performance on this test and college success
 - (4) the test is right about 60 percent of the time in predicting college grades
36. Sus was born July 9, 1948. What will her C.A. be on March 24, 1956?
- (1) 7-7
 - (2) 7-8
 - (3) 7-9
 - (4) 8-0
37. The accomplishment or achievement quotient is defined as the ratio of the educational age to the mental age. Which of the following would be the correct formula for this quotient?
- (1) $A.Q. = \frac{M.A.}{E.A.}$
 - (2) $A.Q. = \frac{E.A.}{M.A.}$
 - (3) $A.Q. = E.A. \times M.A.$
 - (4) none of the above
38. The I.Q. is the ratio of the M.A. to the C.A. times 100. If you know the C.A. and the I.Q., which of the following formulas would you use to find the M.A.?
- (1) $M.A. = \frac{I.Q.}{C.A.} \times 100$
 - (2) $M.A. = \frac{C.A.}{I.Q.} \times 100$
 - (3) $M.A. = \frac{I.Q.}{C.A. \times 100}$
 - (4) $M.A. = \frac{I.Q. \times C.A.}{100}$
39. An appropriate test is said to have curricular validity. Which of the following testing situations would be most likely to have this characteristic?
- (1) a personality test in a physics course
 - (2) a composition test in a literature course
 - (3) repairing a broken tool in a shop mechanics course
 - (4) a test of facts and knowledge in a home economics course
40. A major use of testing is for diagnosis. Which of the following test situations represents the best example of the foregoing statement?
- (1) a comprehensive achievement battery at the end of high school
 - (2) an achievement battery given early in the year
 - (3) an intelligence test
 - (4) a series of tests used to determine a student's grade
41. If Bill scored at the 88th percentile in Social Service on the Kuder Preference Test, it would indicate that:
- (1) Bill got 88% of the answers correct
 - (2) he has more ability in Social Service than 88% of his norm group
 - (3) only 12% of the norm group showed more interest in Social Service than he did
 - (4) that 88 out of 100 will do better than he did on this test

Parts III and IV - Interpretation and Extrapolation

Data are given below on five pupils enrolled in a class of 30 ninth graders. The test data are based on performance at the end of the first semester. Read over the summary and then show to which pupil each statement best fits by marking the pupil's number on the answer sheet. (Numbers 42 through 46)

| <u>Pupil</u> | <u>I.Q.</u> | <u>Calif. Ach. Test Performance</u> | | | <u>Teacher's estimate of Ach. Rank in Class</u> |
|--------------|-------------|-------------------------------------|--------------|--------------|---|
| | | <u>Arith.</u> | <u>Read.</u> | <u>Lang.</u> | |
| | | 1 | 86 | 9.1 | |
| 2 | 99 | 9.7 | 9.6 | 9.5 | 14 |
| 3 | 132 | 9.5 | 9.8 | 10.2 | 12 |
| 4 | 138 | 11.8 | 12.3 | 12.0 | 3 |
| 5 | 101 | 10.0 | 10.1 | 10.9 | 4 |

42. The pupil who should be doing considerably better in his school achievement.
43. The accuracy of the I.Q. seems most doubtful in which case?
44. A bright student making good use of his ability.
45. Teacher regards abilities too highly according to test results.
46. Teacher's rank most consistent with test scores.

The five students for whom the data are given below are in kindergarten. These test data are based on test performance at the beginning of the second semester. After examining the data indicate which pupil best fits each of the following statements by marking the number of the student on the answer sheet. (Numbers 47-53)

| <u>Student</u> | <u>C.A.</u> | <u>I.I.A. on Stanford Binet</u> | <u>Percentile Rank on Readiness Test</u> |
|----------------|-------------|---------------------------------|--|
| 1 | 5-10 | 7-4 | 72 |
| 2 | 6-4 | 5-4 | 22 |
| 3 | 5-10 | 5-5 | 64 |
| 4 | 5-8 | 5-6 | 45 |
| 5 | 5-6 | 6-10 | 38 |

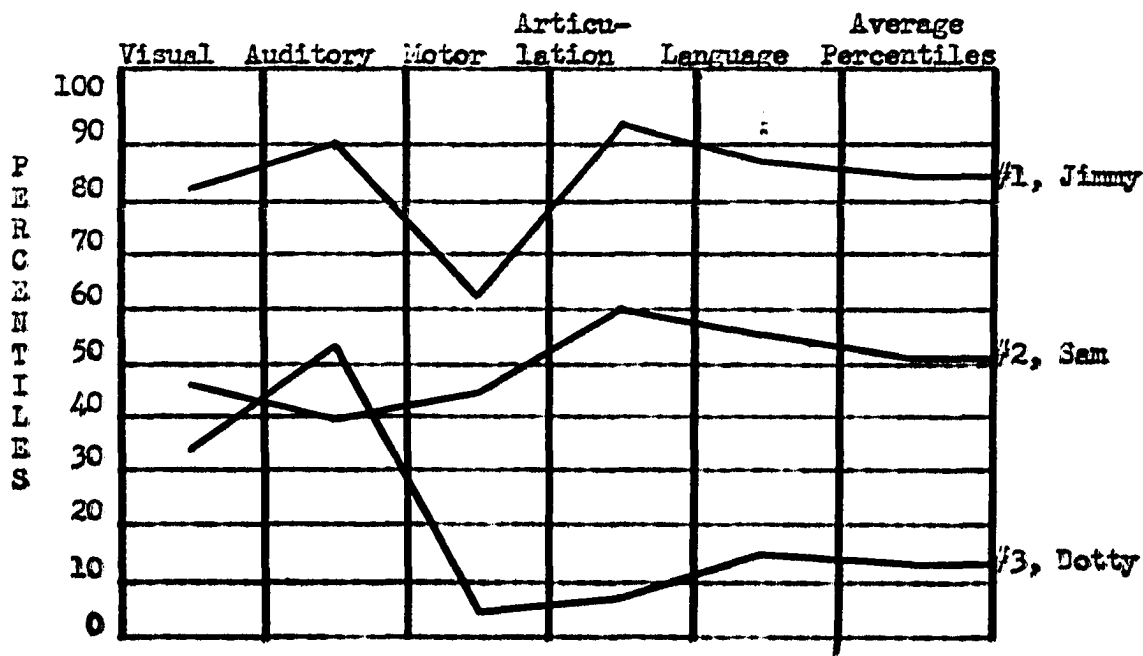
Which student:

47. Is most ready at present for first grade work?
48. Is apparently in need of stimulating experiences but has fairly high aptitude?
49. Will probably not be able to keep up with the average in the first grade?
50. Apparently comes from a very stimulating environment?
51. Is most characteristic of the average for this group?
52. Can you predict will have the lowest ability three years from this time?
53. Is least ready for first grade work?

Examine the data and profiles of these three entering first graders, and answer questions that follow. (Numbers 54 to 60)

1. Jimmy, age 6-0, I.Q. 7-4
2. Sam, age 6-1, I.Q. 7-5
3. Doty, age 6-3, I.Q. 5-3

Reading Readiness Abilities



On your answer sheet mark

- (1) if the statement is most true of #1, Jimmy
- (2) if the statement is most true of #2, Sam
- (3) if the statement is most true of #3, Doty
- (4) if the statement does not validly apply to any of the three students

54. Is at the kindergarten level mentally.
55. Requisites best developed for participation in beginning reading experiences.
56. Teacher will most likely provide at once activities and exercises to develop latent motor ability.
57. Readiness scores most out of line with mental ability.
58. Should have immediate examination by ear specialist.
59. Poorly developed abilities for first grade activities.
60. Apparently very well adjusted socially.

Reading

| | | | | | |
|--|---------------------------------|-------------------|-------------------------|---------------------|---------------------------------|
| | Two yrs. Behind or more | One yr. Behind | Normal | One yr. Advanced | Two yrs. Advanced or more |
| A R I T H M E T I C | Two yrs. Advanced or more | | XX | X | XX |
| | One yr. Advanced | X | XXX | XXX | XXX |
| | Normal | XXX | XXXX XXXX | XXX | |
| | One yr. Behind | XXX | XXX | X | |
| | Two yrs. Behind or more | XX | X | XX | |
| | Average and Retarded Readers | | | Superior Readers | |

(taken from
Cronbach,
Lee J.,
Essentials
of Psycholo-
gical Test-
ing)

Making your judgment on the basis of the information given in the graph, classify each of the following by marking: (Numbers 61 to 67)

- (1) if the item is definitely true
- (2) if the item is probably true
- (3) if the information given is insufficient to make a judgment regarding the truth or falsity of the item
- (4) if the item is probably false
- (5) if the item is definitely false

- 61. In general, children who are good readers will be good in arithmetic.
- 62. If we found a student that was advanced in reading we could, for most purposes, conclude that he would be good in all subjects.
- 63. Of the group of superior readers there are four who are retarded in arithmetic.
- 64. There is a higher relationship between reading and arithmetic than there is between reading and other subjects.
- 65. The cluster of eight students in the center of the graph implies that ability is the same from one subject to the next.
- 66. There are 22 students considered on this graph.
- 67. There is a greater relationship between reading and arithmetic ability among superior readers than among average or retarded readers.

Mr. Tuttle found that his norms did not go high enough to interpret the test score of one of his students. The last four norms are shown below but Sally got a score of 125.

| Score | Grade Equivalent |
|-------|------------------|
| 120 | 12-5 |
| 115 | 12-2 |
| 110 | 12-0 |
| 105 | 11-6 |

68. What would the best estimate of Sally's grade equivalent be?
- (1) 12-6
 - (2) 13-0
 - (3) 13-2
 - (4) over 12-5

Susan's record shows that she has taken two achievement batteries, one at the beginning of the eighth grade and the other in the middle of the tenth grade. Examine their results and answer the questions following. (Numbers 69 to 75)

8th grade, California Achievement Tests

| | <u>Grade</u> | <u>Placement</u> |
|---------------------|--------------|------------------|
| Reading Vocabulary | 9.3 | |
| Read. Comprehension | 7.6 | |
| Total Reading | 8.5 | |
| Arith. Reasoning | 7.0 | |
| Arith. Fundamentals | 9.0 | |
| Total Arithmetic | 8.2 | |
| Language | 8.6 | |
| Total for Battery | 8.7 | |

10th grade, Iowa Tests of Ed. Development

| | <u>Percentile</u> | <u>Score</u> |
|--|-------------------|--------------|
| Understanding of basic social concepts | | 58 |
| Background of natural science | | 45 |
| Correctness of writing | | 73 |
| Quantitative thinking | | 71 |
| Ability to interpret reading materials in the natural sciences | | 46 |
| Ability to interpret reading materials in the social studies | | 51 |
| Ability to interpret literary materials | | 62 |
| General Vocabulary | | 74 |
| Uses of sources of information | | 78 |

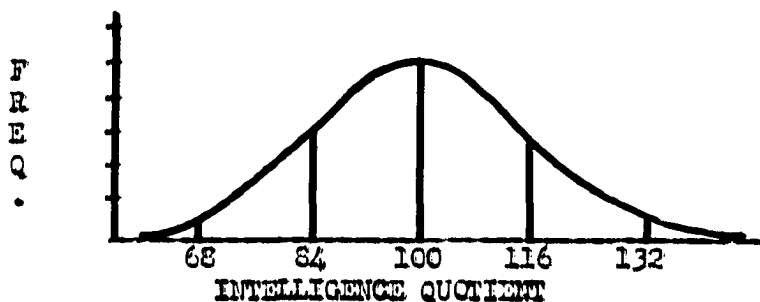
69. There is most disagreement of the two test batteries in:
- (1) ability to think quantitatively
 - (2) knowledge of fundamentals
 - (3) vocabulary
 - (4) ability to read and interpret reading
70. If we were to administer a test such as the California Test of English Usage at the end of the first semester in grade 11, what approximate grade placement would you predict from the previous scores?
- | | |
|----------|----------|
| (1) 10.0 | (3) 12.0 |
| (2) 11.0 | (4) 13.0 |
71. We might infer that Susan came from what type of home environment?
- (1) unstimulating and deprived atmosphere
 - (2) a home that emphasized deep thinking and problem solving
 - (3) a stimulating environment
 - (4) a good home physically, but with parents that didn't care about schooling
72. There is most agreement of the two test batteries in:
- (1) mathematics
 - (2) reading comprehension
 - (3) scientific knowledge
 - (4) vocabulary
73. Which general ability seems to be the strongest? Ability to:
- (1) deal with specifics
 - (2) interpret and solve problems
 - (3) reason and understand
 - (4) draw conclusions
74. On the basis of these test results how well would you expect Susan to do in a course in high school physics?
- (1) below average
 - (2) average
 - (3) above average
 - (4) very well
75. How does the over-all test performance of the Iowa Test compare with the Progressive Test?
- (1) much lower
 - (2) lower
 - (3) about the same
 - (4) higher

Examine the following table and answer questions, numbers 76-79, on the basis of the data alone.

| Mental Age Range by School Grade | |
|----------------------------------|---------------------------------------|
| Grade | M.A. Range (2nd. to 98th. percentile) |
| 11 | 8.4 |
| 9 | 8.4 |
| 7 | 7.2 |
| 5 | 5.6 |
| 3 | 4.8 |
| 1 | 3.6 |

76. A teacher would have her greatest problem of individual differences at what grade level?
 (1) third (3) seventh
 (2) fifth (4) ninth
77. What would be your best estimate of the M.A. range in grade 6?
 (1) 6.2 (3) 6.6
 (2) 6.4 (4) 6.8
78. A high school teacher will find differences between the extremes of the mental ages of approximately:
 (1) 4 to 6 years (3) 8 to 10 years
 (2) 6 to 8 years (4) 10 to 12 years
79. What would be your best estimate to the nearest year of the M.A. range in kindergarten?
 (1) 2 years (3) 4 years
 (2) 3 years (4) 5 years

Study the curve given and answer the questions following. Nos. 80 to 85



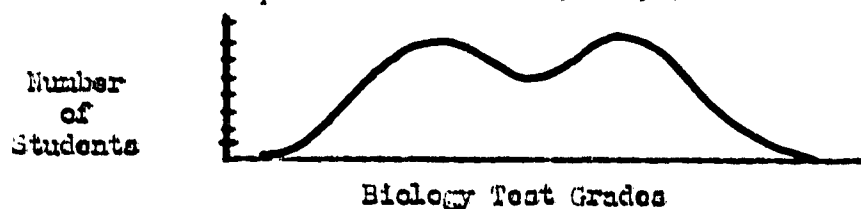
80. The number of people getting I.Q.'s of 140 would be equal to the number getting I.Q.'s of:
 (1) 60 (3) 80
 (2) 68 (4) 132
81. You would expect approximately what percent of the people to have I.Q.'s of less than 50?
 (1) .5% (3) 5%
 (2) 2% (4) 10%
82. According to this curve which in the most common score among the following?
 (1) 84 (3) 86
 (2) 130 (4) 112
83. We would expect that about 95-100 of the I.Q.'s would fall below:
 (1) 100 (3) 116
 (2) 110 (4) 132
84. The top I.Q. according to this curve would be:
 (1) 132 (3) 150
 (2) 140 (4) impossible to tell
85. The greatest number of people would fall in which of the following I.Q. ranges?
 (1) 68-84 (3) 100-116
 (2) 84-92 (4) 116 and above

The following test scores are available on Tom, a senior in high school. From these data answer the questions that follow. (Numbers 86 to 88)

| | |
|--|-----------------------|
| Terman-McLanar Test of Mental Ability - Age 15; I.Q. 143 | |
| Kuder Preference Record - <u>Significantly high</u> | <u>Definitely low</u> |
| Computational | Social Science |
| Scientific | Clerical |
| Literary | |
| Heston Personal Adjustment Inventory - Senior Norms | |
| Analytical Thinking | 96 |
| Home Satisfaction | 70 |
| Emotional Stability | 60 |
| Sociability | 8 |
| Confidence | 12 |
| Personal Relations | 6 |

86. Tom's ability is best described as:
- (1) above average
 - (2) superior
 - (3) very superior
 - (4) high genius
87. Tom's scores indicate that he would be best suited for:
- (1) research
 - (2) medicine
 - (3) teaching
 - (4) law
88. Tom's score patterns indicate a need to:
- (1) widen his scope of interest
 - (2) see a psychiatrist
 - (3) set a definite goal
 - (4) improve his sociability

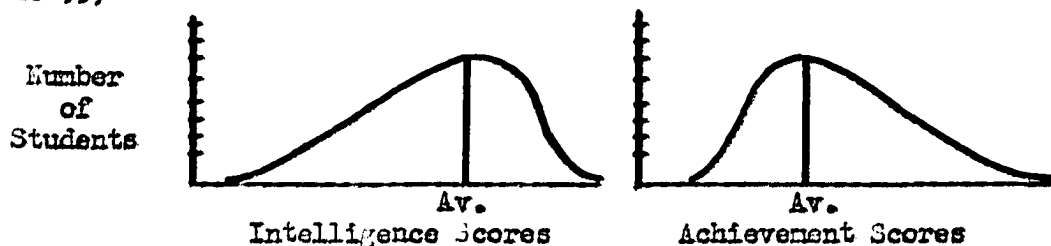
Mr. Smith gave a biology test in his class, a typical sophomore group. He drew a curve showing the distribution of the test scores. Refer to this curve to answer questions numbers 89 to 91.



89. We would expect to find that the test scores indicated:
- (1) about the same number of high and low grades
 - (2) more high than low grades
 - (3) more low than high grades
 - (4) most of the grades around the average
90. When Mr. Smith assigned grades, he would likely have:
- (1) more A's than F's
 - (2) more D's than C's
 - (3) more C's than B's
 - (4) more C's than D's plus B's

91. The best guess we could make about Mr. Smith's students with regard to time the students studied for the test is:
- (1) they all studied for the test
 - (2) some studied and some didn't but most of them did
 - (3) many studied and many didn't study
 - (4) the average student studied pretty hard

Study the curves given and answer the questions following. (Numbers 92 to 95)



Curve #1 shows the distribution of I.Q.'s in a class
 Curve #2 shows the distribution of total achievement test scores in the same class

92. The class, as a group, appears to be:
- (1) underachievers
 - (2) overachievers
 - (3) achieving in line with their ability
 - (4) it is impossible to determine their rate of achievement
93. A student in this class who had an average I.Q. would probably have:
- (1) average achievement
 - (2) slightly above average achievement
 - (3) high achievement
 - (4) slightly below average achievement
94. The range of intelligence test scores as compared to the range of achievement test scores:
- (1) tends to be greater
 - (2) tends to be less
 - (3) can't tell from the data given
 - (4) tends to be about the same
95. A best guess with regard to the students of this class would be that they are:
- (1) having too many assignments
 - (2) not motivated to do school work
 - (3) doing about what we would expect
 - (4) having too much social activity

APPENDIX D

May 23, 1957

Dear Student:

During the fall semester while you were enrolled in Educational Psychology 62, some of your instructors started a research project which we think will help us improve the teaching of the unit on tests and measurements in this course. You were given a test over this unit as a pretest and as a final after completing the unit. We would like for you to take this test once more—four months after the course—to help us find the things you remember best from this unit.

We realize that you are very busy right now, so we are taking this means of letting you take the examination at your own convenience. Please complete the enclosed examination using the answer sheet provided and return both the test and the answer sheet in the enclosed self-addressed envelope.

We sincerely appreciate your cooperation.

Thank you,

Charles O. Neidt
Chairman, Department of
Educational Psychology
and Measurements

CGI:kat

Frequencies of Wrong-Right-Wrong and Wrong-Right-Right Responses to Each Item on the Successive Administrations of the Tests

| Item No. | Wrong-Right-Wrong | Wrong-Right-Right | Item No. | Wrong-Right-Wrong | Wrong-Right-Right | Item No. | Wrong-Right-Wrong | Wrong-Right-Right |
|----------|-------------------|-------------------|----------|-------------------|-------------------|----------|-------------------|-------------------|
| 1 | 16 | 20 | 36 | 28 | 17 | 71 | 23 | 21 |
| 2 | 8 | 8 | 37 | 15 | 30 | 72 | 20 | 32 |
| 3 | 17 | 24 | 38 | 30 | 30 | 73 | 24 | 30 |
| 4 | 3 | 15 | 39 | 14 | 30 | 74 | 26 | 14 |
| 5 | 14 | 36 | 40 | 19 | 27 | 75 | 29 | 34 |
| 6 | 7 | 33 | 41 | 12 | 31 | 76 | 23 | 31 |
| 7 | 15 | 10 | 42 | 3 | 31 | 77 | 12 | 20 |
| 8 | 15 | 24 | 43 | 20 | 27 | 78 | 16 | 25 |
| 9 | 14 | 28 | 44 | 5 | 29 | 79 | 12 | 24 |
| 10 | 11 | 38 | 45 | 16 | 20 | 80 | 3 | 13 |
| 11 | 19 | 29 | 46 | 33 | 23 | 81 | 16 | 23 |
| 12 | 19 | 40 | 47 | 3 | 20 | 82 | 1 | 25 |
| 13 | 18 | 20 | 48 | 28 | 34 | 83 | 23 | 35 |
| 14 | 26 | 37 | 49 | 2 | 20 | 84 | 10 | 18 |
| 15 | 1 | 17 | 50 | 14 | 7 | 85 | 5 | 20 |
| 16 | 10 | 23 | 51 | 9 | 26 | 86 | 37 | 23 |
| 17 | 11 | 26 | 52 | 3 | 26 | 87 | 4 | 16 |
| 18 | 15 | 34 | 53 | 3 | 15 | 88 | 7 | 28 |
| 19 | 57 | 62 | 54 | 7 | 22 | 89 | 13 | 31 |
| 20 | 28 | 54 | 55 | 4 | 25 | 90 | 16 | 23 |
| 21 | 18 | 18 | 56 | 9 | 12 | 91 | 17 | 32 |
| 22 | 32 | 44 | 57 | 19 | 27 | 92 | 14 | 31 |
| 23 | 11 | 20 | 58 | 15 | 6 | 93 | 15 | 28 |
| 24 | 23 | 15 | 59 | 4 | 19 | 94 | 15 | 17 |
| 25 | 12 | 32 | 60 | 24 | 12 | 95 | 12 | 33 |
| 26 | 17 | 25 | 61 | 14 | 24 | | | |
| 27 | 5 | 37 | 62 | 29 | 26 | | | |
| 28 | 7 | 20 | 63 | 26 | 31 | | | |
| 29 | 2 | 16 | 64 | 18 | 28 | | | |
| 30 | 18 | 17 | 65 | 15 | 13 | | | |
| 31 | Eliminated | | 66 | 14 | 11 | | | |
| 32 | 12 | 21 | 67 | 25 | 15 | | | |
| 33 | 7 | 3 | 68 | 18 | 27 | | | |
| 34 | 8 | 0 | 69 | 32 | 17 | | | |
| 35 | Eliminated | | 70 | 36 | 21 | | | |

Frequencies of Wrong-Right-Wrong and Wrong-Right-Right
Responses to Each Item on the Successive Administrations
of the Tests With Chance Occurrence Subtracted

| Item No. | Wrong-Right- Wrong | Wrong-Right- Right | Item No. | Wrong-Right- Wrong | Wrong-Right- Right | Item No. | Wrong-Right- Wrong | Wrong-Right- Right |
|----------|-----------------------|-----------------------|----------|-----------------------|-----------------------|----------|-----------------------|-----------------------|
| 1 | -- | 12 | 36 | 4 | 9 | 71 | -- | 13 |
| 2 | -- | -- | 37 | -- | 22 | 72 | -- | 24 |
| 3 | -- | 16 | 38 | 6 | 22 | 73 | -- | 22 |
| 4 | -- | 7 | 39 | -- | 22 | 74 | 2 | 6 |
| 5 | -- | 28 | 40 | -- | 19 | 75 | 5 | 26 |
| 6 | -- | 25 | 41 | -- | 23 | 76 | -- | 23 |
| 7 | -- | 2 | 42 | -- | 25 | 77 | -- | 12 |
| 8 | -- | 16 | 43 | -- | 21 | 78 | -- | 17 |
| 9 | -- | 20 | 44 | -- | 23 | 79 | -- | 16 |
| 10 | -- | 30 | 45 | -- | 14 | 80 | -- | 5 |
| 11 | -- | 21 | 46 | 11 | 17 | 81 | -- | 15 |
| 12 | -- | 32 | 47 | -- | 14 | 82 | -- | 17 |
| 13 | -- | 12 | 48 | 6 | 28 | 83 | -- | 27 |
| 14 | 2 | 29 | 49 | -- | 14 | 84 | -- | 10 |
| 15 | -- | 9 | 50 | -- | 1 | 85 | -- | 12 |
| 16 | -- | 15 | 51 | -- | 20 | 86 | 13 | 15 |
| 17 | -- | 18 | 52 | -- | 20 | 87 | -- | 8 |
| 18 | -- | 26 | 53 | -- | 9 | 88 | -- | 20 |
| 19 | 33 | 54 | 54 | -- | 14 | 89 | -- | 23 |
| 20 | 4 | 46 | 55 | -- | 17 | 90 | -- | 15 |
| 21 | -- | 10 | 56 | -- | 4 | 91 | -- | 24 |
| 22 | 8 | 36 | 57 | -- | 19 | 92 | -- | 23 |
| 23 | -- | 12 | 58 | -- | -- | 93 | -- | 20 |
| 24 | -- | 7 | 59 | -- | 11 | 94 | -- | 9 |
| 25 | -- | 26 | 60 | -- | 4 | 95 | -- | 25 |
| 26 | -- | 19 | 61 | -- | 18 | | | |
| 27 | -- | 31 | 62 | 7 | 20 | | | |
| 28 | -- | 14 | 63 | 4 | 25 | | | |
| 29 | -- | 10 | 64 | -- | 22 | | | |
| 30 | -- | 11 | 65 | -- | 7 | | | |
| 31 | Eliminated | | 66 | -- | 5 | | | |
| 32 | -- | 15 | 67 | 3 | 9 | | | |
| 33 | -- | -- | 68 | -- | 19 | | | |
| 34 | -- | -- | 69 | 8 | 9 | | | |
| 35 | Eliminated | | 70 | 12 | 13 | | | |