AN INVESTIGATION OF THE EFFECTS OF DISCOVERY LEARNING ON RETENTION AT TWO LEVELS OF MENTAL FUNCTIONING

A Thesis Presented to

the Faculty of the College of Education

The University of British Columbia

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

by

Leonard Paul Kroeker

August 1967

We accept this thesis as conforming to the

required standard

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and Study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Education

The University of British Columbia
Vancouver 8, Canada

Date August 15, 1967.

ABSTRACT

It was hypothesized that greater mean retention would occur at the Knowledge and Application levels of mental functioning, as defined by Bloom's Taxonomy of Educational Objectives, among students taught by a discovery method than among students taught by a more conventional, lecture-demonstration method. It was also hypothesized that greater mean retention would occur at the Application level than at the Knowledge level among students taught by a discovery method than among students taught by a lecture-demonstration method.

Each of two ninth grade science classes in a single school was taught a heat unit using one of the methods mentioned above. The teaching methods were assigned randomly to intact classes, both handled by the same teacher.

Two multiple choice achievement tests covering the content of the heat unit were constructed; one consisting of items in the Knowledge category and the other consisting of items in the Application category. A tryout of each of these tests was conducted upon 160 students in a single school thus allowing the elimination of unsuitable items, specifically, items whose discrimination indices were negative and those whose difficulty indices were either too high or too low. The resulting unit tests, with forty Knowledge and thirty two Application items respectively, were administered to the students of the two classes both immediately following and six weeks following the conclusion of the heat unit.

The reliability coefficients of the Knowledge and Application tests, estimated by correlating the half test scores and applying the Spearman-Brown formula, were .82 and .80 respectively.

Covariates, chosen on the basis of their correlation with loss scores (measures of retention), were used to adjust experimental and control group loss score means. The analysis of covariance showed a significant difference between Application loss score means and a non-significant difference between Knowledge loss score means at the pre-set 5 percent significance level. It was therefore concluded that this experiment provided evidence for the acceptance of the experimental hypothesis dealing with the retention of application objectives and for the rejection of the experimental hypothesis dealing with the retention of the knowledge objectives.

Items, matched on the basis of content, difficulty index, and discrimination index were selected from the two unit tests to form Knowledge and Application subtests. Loss scores were calculated from the subtest results and the differential retention hypothesis was tested using a Z statistic. The analysis revealed a non-significant difference between differential loss score means at the pre-set 5 percent significance level.

It was concluded that there was no statistical basis to support the hypothesis that greater mean differential retention (between Application and Knowledge) would occur among students taught by a discovery method than among students taught by a lecture-demonstration method.

It was, however, suggested that further experimental refinements might possibly produce significant results when testing the differential retention hypothesis in a future replication of the study. Theoretical and practical implications of the findings were also discussed.

ACKNOWLEDGEMENT

I wish to express my thanks to Prof. H. Cannon of the Faculty of Education for his patient guidance throughout. I am also indebted to Dr. D. McKie of the Faculty of Education for his time spent in discussion and for his many helpful suggestions.

Thanks are also due to Mrs. J. Woodrow of the Faculty of Education for her careful proof-reading and many suggestions regarding the writing of the thesis, and to the late Mr. D. Webster of the Faculty of Education for his careful analysis of the tests used.

Finally, I wish to thank my wife, Mrs. A. Kroeker, for her careful typing of the manuscript.

TABLE OF CONTENTS

CHAPTER		PAGE
I.	THE PROBLEM AND HYPOTHESIS	1
	1.0 Introduction	1
	1.1 The Problem	3
	a) Statement of the Problem	3
	b) Statement of the hypotheses	4
	c) The null hypotheses	4
	d) The alternative hypotheses	5
	1.2 Definition of Terms	5
	1.3 Discussion of the Problem	11
	1.4 Importance of the Study	13
II.	REVIEW OF THE LITERATURE	16
	2.0 Introduction	16
	2.1 Literature Related to Teaching Methods	16
	2.2 Literature Related to Instructional Objectives and Retention	19
	2.3 Summary	23
III.	SPECIFICATION OF RELEVANT PREPARATIONS	24
	3.0 Introduction	24
	3.1 The Content of the Unit	24
	3.2 The Unit Tests	26
	a) Content of the unit tests	26
	b) Type of test	27
	3.3 The Pretests	28
	3.4 The Subjects	32
	3.5 The Tryout Problem	33
	3.6 The Experimental Design	35

CHAPTER PAGE

IV. METHODOLOGY 38

4.0 Introduction 38

4.1 Preparation of the Students 38

4.2 The Instruction of the Unit 40

4.3 Preparation of the Tryout Tests 41

4.4 Administration of the Tryout Tests 43

a) The problem of guessing 43

b) The results of the tryout tests 44

4.5 The Administration of the Pretests 48

4.6 The Preparation and Administration of the Final Test 48

a) The selection of the items 48

b) The construction of the unit tests..... 50

c) The administration of the unit tests 51

4.7 The Administration of the Retest 53

a) Administration of the retest 53

b) The test results 53

V. THE STATISTICAL ANALYSIS 55

5.0 Introduction 55

5.1 The Unit Test Results and Reliabilities 55

a) The unit test scores 55

b) The reliability of the unit tests 59

5.2 The Pretest Results 60

5.3 The Statistical Test of Hypothesis I 63

5.4 The Statistical Test of Hypothesis II 65

5.5 The Loss Score Reliabilities 67

5.6 The Subtests Used in the Test of Hypothesis III 68

CHAPTER PAGE

5.7 The Statistical Test of Hypothesis III 71

VI. CONCLUSIONS AND SUMMARY 73

6.0 Introduction 73

6.1 Limitations of the Study 73

6.2 Conclusions 74

6.3 Implications of the Study 75

6.4 Possibilities for Further Research 76

6.5 Summary of the Investigation 76

BIBLIOGRAPHY 79

APPENDIX A. Test Directions 83

APPENDIX B. Sample Lessons 85

APPENDIX C. The Unit Tests 90

APPENDIX D. The Data 104

APPENDIX E. Derivation of Test Statistic 108

LIST OF TABLES

TABLE		PAGE
I.	Distribution of the Item Difficulty Indices for the Knowledge and Application Tryout Tests	46
II.	Distribution of the Item Discrimination Indices for the Knowledge and Application Tryout Tests	47
III.	Distribution of Items by Content Category in the Knowledge and Application Unit Tests	52
IV.	Distribution of the Unit Test Scores	56
V.	Distribution of the Difficulty and Discrimination Indices for the Unit Tests	57
VI.	Distribution of the Loss Scores	58
VII.	Distribution of the Verbal Reasoning and Numerical Ability Pretest Scores	61
VIII.	Distribution of the Read General Science Pretest Scores	62
IX.	Distribution of Items by Content Category in the Knowledge and Application Subtests	69
X.	Distribution of Difficulty Indices in the Subtests	70

CHAPTER I

1.0 INTRODUCTION

The old high school general science courses in British Columbia although intended to achieve major objectives such as the development of fundamental scientific concepts, the development of understanding of scientific principles and the acquisition of an appreciation of the scientific method, have in fact tended toward an overemphasis upon factual recall for examination purposes.[1]

The advent of the P.S.S.C., B.S.C.S., and the CHEM. STUDY courses in the senior high school program has issued a challenge to the old general science courses in that the former may tend to achieve common desirable objectives such as those listed above in a significantly greater way than the latter. It is believed, for example, by the authors of the new science courses listed above that intuitive concept development is given greater opportunity to flourish under conditions of empirical laboratory investigation and that these concepts so developed lead to a greater understanding of the inter-relationships within the structure of the subject.

As a result of the concern whether high school general science courses met objectives such as those listed above, science revision committees in British Columbia have been prompted to write alternative science units which have recently been collected and adopted as the prescribed general science courses in British Columbia high schools by

---

[1]British Columbia Department of Education, Science Curriculum Bulletin, (Victoria, B.C.: The Queen's Printer, 1963) p. 1.

the Curriculum Division of the Department of Education.

The physics units in these general science courses for grades eight, nine and ten are short physics courses at a very elementary level. The units, developed within the framework of objectives alluded to above, have been designed to allow students to discover basic physical principles for themselves.[2]

The basis for the implementation of these courses has been the belief that self discovery learning will achieve the above objectives more effectively than other instructional methods. The extent to which this belief is justifiable is a problem worthy of investigation.

Whether or not students will discover principles for themselves will depend upon, among other things, the materials and facilities available, the time allotted, and the preparation, industriousness, and resourcefulness of the teacher. Therefore, any investigation central to the problem of self discovery in the classroom situation requires information concerning the factors mentioned above.

It is within the context of the above discussion that the present study finds its roots. A single aspect of discovery learning in the classroom will be investigated in order to shed some light on the efficiency of self discovery classroom learning compared to lecture demonstration learning.

---

[2]British Columbia Department of Education, Science Curriculum Bulletin, (Victoria, B.C.: The Queen's Printer, 1965), p. 1.

1.1 THE PROBLEM

a) Statement of the problem. The purpose of this experiment is to compare the effectiveness of a laboratory centered method of science teaching and a lecture demonstration method, using retention as the criterion variable at two levels of mental functioning as defined by Bloom's Taxonomy of Educational Objectives.[3]

A unit on heat will be taught to two classes of grade nine students. The first class will be taught through self discovery and the second through a lecture demonstration approach.

Retention of the material in the heat unit will be measured by the difference in an individual's scores on the same test given on two occasions separated by about one and a half months.

Two self-defining achievement tests will be constructed in order to test two separate levels of mental functioning; as defined by the Taxonomy of Educational Objectives[4] they are the knowledge level and the application level.

Each of these two tests will then be used separately in the determination of retention on the Knowledge criterion and on the Application criterion. Each separate set of test results will be compared to determine whether retention has been affected by the teaching method used.

---

[3]Benjamin S. Bloom, et al., Taxonomy of Educational Objectives, Handbook I: Cognitive Domain (New York: Longmans, Green & Co., 1956).

[4]Ibid.

A secondary investigation will also be carried out in order to determine whether teaching method differentially affects retention on the Knowledge criterion and retention on the Application criterion.

b) Statement of the hypotheses. It is hypothesized that there is significantly greater mean retention at the Knowledge level of mental functioning among ninth grade science students taught by means of a self discovery method than among those taught by means of a lecture demonstration method.

Similarly, it is hypothesized that there is significantly greater mean retention at the Application level of mental functioning among ninth grade science students taught by means of a self discovery method than among those taught by means of a lecture demonstration method.

Finally, it is hypothesized that there is significantly greater mean differential retention at the Application level than at the Knowledge level of mental functioning among ninth grade science students taught by means of a self discovery method than among those taught by means of a lecture demonstration method.

c) The null hypotheses.

1. There is no significant difference in mean loss score at the Knowledge level between the experimental group and the control group when the significance level is set at five percent.
2. There is no significant difference in mean loss score at the Application level between the experimental group and the control group when the significance is set at five percent.

3. There is no significant difference between the difference in experimental and control group loss score means at the Application level and the difference in experimental and control group loss score means at the Knowledge level when the significance level is set at five percent.

d) The alternative hypotheses.

1. The experimental group mean loss score is significantly smaller than the control group mean loss score at the Knowledge level when the significance level is set at five percent.

2. The experimental group mean loss score is significantly smaller than the control group mean loss score at the Application level when the significance level is set at five percent.

3. The difference in experimental and control group loss score means at the Application level is significantly greater than the difference in experimental and control group loss score means at the Knowledge level when the significance level is set at five percent.

1.2 DEFINITION OF TERMS

Knowledge level. According to Bloom, knowledge "involves the recall of specifics and universals, the recall of methods and processes, or the recall of a pattern, structure, or setting..... The knowledge objectives emphasize most the psychological processes of remembering."[5]

---

[5]Bloom, op. cit., p. 201.

It is this definition which will be referred to when the phrase "knowledge level" is used in this study.

Application level. According to Bloom, application refers to the use of an appropriate abstraction in the solution of a problem in which no mode of solution has been specified.[6] The abstraction referred to may be used in particular and concrete situations. They "may be in the form of general ideas, rules of procedures, or generalized methods..... and may also be technical principles, ideas, and theories which must be remembered and applied."[7] The behaviors thus specified will be referred to as the "application level" in this study.

Academic-technical stream. The academic-technical stream refers to the British Columbia high school stream composed of students who are interested primarily in furthering their education at the university level.

Experimental group. The experimental group refers to the ninth grade science students in this study who were taught by means of a self discovery method.

Control group. The control group refers to the ninth grade science students in this study who were taught by means of a lecture demonstration method.

---

[6]Bloom, op. cit., p. 120.

[7]Bloom, op. cit., p. 205.

Loss score. Loss score refers to the signed difference obtained when an individual subject's second score is subtracted from his first score on the same test.

Mean loss score. Mean loss score refers to the average of the signed differences of all the subjects within the same group.

Mean retention. Mean retention is defined as the measure which is the average score of all the subjects within a single group on the second administration of the test to the group. Mean retention and mean loss score are complementary in that their sum is the average score of the same subjects on the first administration of the test.

Self discovery method. Self discovery method is defined as the procedure in which students are guided by the teacher in discovering for themselves physical principles in the laboratory.

The behavior of the teacher will depend upon the goals to be achieved and the present behavior of the students. These goals include ones cited earlier, namely, the development of fundamental scientific concepts, the development of understanding of scientific principles, and the acquisition of an appreciation of the scientific method. Teaching procedures which may lead to these goals are the following:

i) the opportunity for each student to proceed at his own rate;

ii) the provision for a basic learning sequence for all students;

iii) the provision for a laboratory experience as an integral part of the learning sequence; and

iv) the provision for enrichment through the investigation of optional problems.

Student understanding of the scientific method of inquiry is brought about through emphasis on careful observation, diligent record keeping, critical evaluation of assumptions and circumspect procedure in the drawing of inferences.

The teacher is to define the task for the students by requesting they prepare at home for the experiment to be performed by reading carefully the distributed mimeographed experiment description and guide questions and by discussing with them on the day the experiment is to be performed the nature of the problem under investigation, the possible hypotheses and methods of investigation, and the operation of any unfamiliar equipment. The teacher is not to demonstrate how the experiment is to be performed however.

The teacher is to be well organized and have all materials ready for distribution when the students enter the classroom. He is expected to communicate well with his students and manage his classroom effectively in terms of productive student behavior and time allotted to various activities. The teacher is expected to maintain a friendly atmosphere and to draw on his own resourcefulness and creativity to promote intrinsic student motivation by preparing additional materials for optional investigation. He is to give only sufficient information to kindle interest and is not to solve a problem for the student.

The students are to be grouped in threes and are to perform the experiment, answer the guide questions, and formulate their own conclusions. They are to submit the experimental report in a format of their own choice at the end of the science period and they will receive the graded report at the beginning of the following science period. The grading procedure is calculated to encourage the students through

constructive written comments and stimulate additional investigation. A brief discussion concerning the results of the experiment will also take place when the reports are returned. It is expected that some students will be ahead of others; they may therefore continue with the next experiment in the sequence or continue with investigation of interesting problems arising from their previous work.

Lecture demonstration method. Lecture demonstration is defined as the procedure in which students learn through listening to the teacher's verbal presentation of material, through watching his demonstration of physical principles and through completing the written assignments required of them by the teacher.

The chief goals for this method are the same as those for the self discovery method, namely, the development of scientific concepts, the development of understanding of scientific principles, and the acquisition of an appreciation of the scientific method.

During lectures as well as demonstrations the students are encouraged to observe carefully, keep records accurately, evaluate assumptions critically, and draw inferences cautiously.

The teacher is to define the task for the students by presenting aspects of theory both verbally and written on the blackboard in order to illuminate certain physical principles in the material taught. He is to consider problems similar to those posed in the self discovery method. In a similar way he is to discuss the nature of the problem to be investigated by demonstration experiment and the possible hypotheses and methods that might be used in the investigation. The teacher is to perform the demonstration experiment and ask specific questions of

students in order to provide cues for observation. Guide questions similar to those used in the self discovery method are to be answered by the students and finally the students are asked to formulate their own conclusions.

The teacher is similarly expected to exhibit the same characteristics of organization, communication, management, friendliness, and creativity in the performance of his tasks in the lecture demonstration method as are exhibited in the self discovery method. Adequate preparation and anticipation should serve to control these characteristics which have been related to student performance.[8]

Since the teacher may have more time available in the lecture demonstration method because of the expected efficiency in comparison to the self discovery method he may elicit more responses from students to assist in the formulation of concepts and the presentation of ideas. Additional demonstrations pertinent to the development of the unit may also be performed.

The students are to complete demonstration experiment reports in a given format and are to submit the reports and guide questions at the end of a science period in which a demonstration experiment has been performed. These are to be graded in the same manner as those submitted by students in the self discovery method and are also to be returned at the beginning of the following science period with a full discussion of the experimental results and answered guide questions.

---

[8]Fletcher G. Watson, "Research on Teaching Science," Handbook of Research on Teaching, N. L. Gage, editor (Chicago: Rand McNally & Co., 1963), pp. 1037-1040.

1.3 DISCUSSION OF THE PROBLEM

The question of the relative merits of various teaching methods has been considered by a number of educational researchers. Although the research to date on the effects of teaching method on achievement is inconclusive and it is unlikely that any one method is superior to any other when the over-all effects are appraised, "the best one might hope for would be slight differences in teaching effectiveness within narrow aspects of the learning process, and this is roughly what is found by empirical research."[9]

One of the problems of the researcher in science education is whether or not the new general science courses which have been designed to be taught by discovery methods are necessarily superior to the science courses which have been replaced. This is not the object of the present study. However, the discovery method has such an apparently heavily invested interest in the new science courses it has been proposed in this study to determine the effectiveness of a self discovery teaching method in comparison to a lecture demonstration method which one might reasonably expect to be used for instructional purposes because of inadequate materials or facilities, inadequate teacher preparation in laboratory science or overlarge classes.

In order to clarify this problem a discussion of the problem in context is in order. Traditionally the laboratory in high school science courses has been given a place of secondary importance since

---

[9]Norman E. Wallen and Robert M. W. Travers, "Analysis and Investigation of Teaching Methods," Handbook of Research on Teaching, N. L. Gage, editor (Chicago: Rand McNally & Company, 1963), p. 500.

the laboratory has long been regarded as a location where verification of classroom science learning occurred. Scientists, however, use the laboratory primarily as a place for investigation. If the student is to appreciate how the scientist works and thinks it would seem apparent that the laboratory should serve the needs of science teaching in the same manner.

Many individuals currently involved in science education believe increases in student involvement in science program activities is essential to greater subject matter understanding. This belief is borne out by the replacement of science courses cited earlier. Whether or not this action can be supported by research findings is a question whose solution may in a partial measure be indicated by the results of this study.

Suchman asserts that educators "have been prompted to reformulate their methods to capitalize on the intense motivation and deep insight that seem to accrue from the 'discovery' approach to concept attainment."[10] The Physical Science Study Committee has adopted the view that students can be sufficiently interested in physics by an experimental approach and can thus be motivated to perform the difficult task of scientific discovery. The Department of Education's philosophy underlying the new science courses appears to be similar in this respect.

Although it appears that no single teaching method has been demonstrated to be clearly superior to any other in terms of achievement, the effectiveness of a teaching method seems to be increased when student

---

[10]J. Richard Suchman, "Inquiry Training: Building Skills For Autonomous Discovery," Merrill-Palmer Quarterly, vol. 7 (July, 1961), p. 147.

inquiry is emphasized.[11] Thus, it is this aspect of student inquiry which one might reasonably expect to be stimulated to a greater extent by self discovery learning than by a lecture demonstration method.

In order to form a basis upon which to evaluate the relative merits of the two teaching methods listed above, several criteria other than achievement might be considered. Among these, the most prominent are retention, transfer of training, and attitudes towards science. For the purposes of this study retention at two levels of mental functioning as defined by the Taxonomy has been chosen as the criterion upon which to evaluate the two methods.

1.4 IMPORTANCE OF THE STUDY

It has been stated earlier that the replacement of general science courses has occurred in British Columbia upon limited supportive research evidence. The working hypothesis of the Department of Education has evidently been that a laboratory centered approach to science teaching is more beneficial to students in terms of understanding scientific facts and laws than a non-laboratory centered approach.[12]

This study will attempt to support this working hypothesis by comparing a typical non-laboratory centered approach with a laboratory centered approach to science teaching. As such an attempt, it is expected to provide information concerning only a single portion of the

---

[11]U.S. Department of Health, Education, and Welfare, Office of Education, New Dimensions in Higher Education (Effectiveness in Teaching No. 2. Washington, D. C. : 1960), p. 15.

[12]British Columbia Department of Education, Junior Secondary School Science (Victoria, B. C. : The Queen's Printer, 1966), p. 3.

problem leaving sufficient scope for much further investigation. However, should it prove fruitful, a significant contribution to this field will have been made.

It is apparent from previous research findings that teaching method appears to have a non significant effect on achievement. This does not suggest, however, that further research on teaching methods be abandoned but rather that further research be carried out to devise other teaching methods which perhaps make as much use as possible of a wide range of learning principles.

Also, further investigation is necessary to determine the possibility of the existence of a significant relationship among other instructional objectives and teaching methods. One of these instructional objectives is retention, which of course is the basis for the present study.

Although much investigation concerning retention has been of a general nature, very little has dealt with the nature of specific outcomes. This study is not only attempting to establish retention as a function of teaching method but also retention as a function of behavioral objectives. Research to date has included investigation concerning differential retention among the behavioral objectives defined by Bloom at the knowledge and the comprehension level.[13] The usefulness to educators of research findings indicating significantly greater retention on the upper levels of mental functioning is clearly evident especially when considered with the relatively low relationship between

---

[13]William P. McDougall, "Differential Retention of Course Outcomes in Educational Psychology," Journal of Educational Psychology, 49, No. 2, April 1958, p. 53.

tests of some of the more complex cognitive abilities and skills and measures of higher intelligence.[14]

On a more immediate, utilitarian level, confirmation of the research hypotheses might well be sufficient encouragement for science teachers to use the self discovery method, despite the inconveniences, instead of abandoning it out of sheer frustration.

---

[14]Bloom, op. cit., p. 22.

CHAPTER II

REVIEW OF THE LITERATURE

2.0 INTRODUCTION

This chapter surveys the literature pertaining to relevant instructional methods and objectives and to noteworthy retention studies. It also provides a critical discussion of a study, similar to the present one, investigating the effects of instructional methods on achievement.

2.1 LITERATURE RELATED TO TEACHING METHODS.

Teaching methods based on philosophical tradition, folklore, and personal needs of teachers have frequently been used in instructional method comparison studies without much success. This is perhaps attributable to the fact that very little has been done to develop teaching methods on the basis of scientific knowledge of learning.

Studies comparing teaching methods have also lacked scientific sophistication in that the variables involved have reflected few of the properties of well developed scientific variables. Wallen and Travers state that these variables tend to be intuitively derived rather than empirically derived.[1] The following results should therefore come as no surprise. In the 1930's, many studies were undertaken in which outcomes of education in schools reported to be progressive were compared with those in schools reputed to be more traditional. The independent

---

[1]Wallen and Travers, op. cit., p. 466.

variable in this case was not differences in teaching practices but differences in reputation. Characteristically, these studies have shown negative results.[2]

The Eight-Year Study, a comparison between "progressive" and "traditional" school graduates on the basis of success in college, showed that a positive correlation existed between the degree of success in college and the degree of experimentation in the school. However, several major criticisms regarding the internal validity of the study have been made; first, the lack of subject randomization and its potential consequences such as selection-treatment interaction bias and secondly, the possible confounding of the variable under study by factors such as teacher competence, personality, and enthusiasm.[3]

Although exhaustive designing of the two teaching methods is not contemplated it would seem appropriate for the purposes of this study to specify characteristics of optimum teacher behavior with respect to classroom activities for each of the methods. This would meet the criticism levelled at so many methods comparison studies that these studies essentially compare two unknowns and do not provide for the replication of the study. Artificiality of treatment also has less opportunity to endanger the validity of the study.

The value of laboratory work in high school science courses has also been the subject of a large number of studies. Cunningham has found that many of these early studies have suffered from inadequacy of

---

[2] Ibid., p. 468.

[3] Ibid., p. 472 - 473.

statistical treatment, invalidity of design, unreliability of results, and lack of clarity of objectives.[4]

Recently, Boldt has made a comparison between a group of grade nine and ten students taught by a self discovery method and a similar group taught by a demonstration method on the basis of achievement.[5] He reported a significant difference between the means of the two methods groups at the one percent level of significance. The demonstration group performed at a significantly higher level than the experimental group on a teacher constructed, self defining achievement test.

His findings should be interpreted with considerable caution, however. Most of the students involved lacked experience in the use of the laboratory. Specifically, they lacked sufficient knowledge, skill, and experience in handling the apparatus and in interpreting their data. No follow up discussions were held after a lesson with the result that the students in the experimental group appeared to be in doubt concerning the correctness of their conclusions. The experimental group also appeared to have insufficient time to check their conclusions by repeating parts of the experiment.

No attempt was made to compare the groups on the basis of any other teaching objectives.

---

[4]H. A. Cunningham, "Lecture method versus individual laboratory method in science teaching - A Summary," Science Education, 30 (1946), pp. 70 - 82.

[5]Walter B. Boldt, "Grade Placement of an Experimental Unit in Secondary School Physics" (unpublished Master's thesis, the University of British Columbia, Vancouver, 1963).

2.2 LITERATURE RELATED TO INSTRUCTIONAL OBJECTIVES AND RETENTION

Specification of the outcomes in this experiment is to be made in terms of the broad classifications of educational objectives listed in the Taxonomy of Educational Objectives.[6]

The Taxonomy is to be chosen for use in the study because of its classification of educational objectives in observable, describable, behavioral terms.[7] Detailed descriptions and illustrations of the behaviors listed in each category facilitate classification of test items to be used in this study in terms of the objectives which they are to measure.

In view of the results of studies on the retention value of different outcomes it is felt that the educational objectives known as the Knowledge and Application levels in the Taxonomy are representative of the cognitive skills and abilities one would expect to develop through the instruction of an experimental science unit such as the one to be used in this study.

Tyler's studies show that knowledge of specific information is not a lasting outcome of instruction in comparison with the ability to apply principles to new situations.[8]

Furst notes:

> The performance of learning is likely to be greatly enhanced if the various outcomes have acquired some interrelation in behavior.

---

[6]Bloom, op. cit.,

[7]Ibid., p. 5.

[8]Ralph W. Tyler, "Permanence of Learning," With the Technicians, Journal of Higher Education, IV, (April 1933), pp. 203 - 204.

> Isolated skills and items of information tend to be forgotten rapidly but those aspects of behavior which bear a functional relationship to other aspects have a much greater probability of being called in to use periodically and of being reinforced. It is known from research studies that those learning outcomes which are continually reinforced and progressively developed into more generalized modes of reaction are most likely to survive in the long run.[9]

Tyler correlated tests designed to measure different types of outcomes and found low correlation between scores on information and scores on application.[10] He concluded that since students did not develop corresponding degrees of ability in the course outcomes studied, the outcomes must have been different from each other. Tyler also produced evidence that the more complex objectives were retained somewhat better than the less complex objectives.

Furst notes that an educational implication of Tyler's findings is that it is necessary in teaching to aim explicitly at each of these major objectives, rather than to assume that the development of ability to think arises from the acquisition of information.[11]

A major problem facing educators is that of "making learning experiences of students more lasting."[12] The Taxonomy points out the

---

[9]Edward J. Furst, Effect of the organization of learning experiences upon the organization of learning outcomes," Journal of Experimental Education, 18, (1950), p. 215.

[10]Ralph W. Tyler, "The relation between recall and higher mental processes," Education as Cultivation of the Higher Mental Processes, C. H. Judd, editor (New York: The Macmillan Co., 1936), pp. 6-17.

[11]Edward J. Furst, Constructing Evaluation Instruments (New York: Longmans, Green and Co., 1958), pp. 49 - 50.

[12]William P. McDougall, op. cit. p. 53.

need for further research concerning retention of specific course outcomes in the following quotation:

> For the most part research on the problems in retention, growth and transfer has not been very specific with respect to the particular behavior involved. Thus, we are not usually able to determine 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.[13]

McDougall reports that there was differential retention among the behavioral objectives he measured, namely: (a) knowledge, (b) translation, (c) interpretation, and (d) extrapolation.[14] The tests constructed to measure these four behavioral outcomes were administered as a pre-test before a unit on educational psychology was studied, at the completion of the unit, and after a period of four months had elapsed. "The results of the study of retention indicated that the abilities to interpret and extrapolate were retained to a significantly greater degree than the ability to recall or translate this knowledge from one form to another."[15] About 79 percent of the gains in interpolation and extrapolation abilities were retained after four months. Gains in knowledge and in the ability to translate knowledge were retained to a significantly lesser degree, approximately 73 percent.

McDougall also found that tests constructed to measure certain behavioral outcomes performed separate functions as evaluation devices. He suggested that if multiple course outcomes, in line with the

---

[13]Bloom, op. cit., p. 23.

[14]McDougall, op. cit., p. 59.

[15]Ibid.

instructional objectives were to be achieved, it would be necessary to design evaluation instruments to accomplish these separate functions.[16]

Smeltz investigated a problem concerning the extent to which learnings of high school chemistry were retained one year following the completion of the course.[17] He reported that pupils retained approximately 68 percent of the information achieved during the year of instruction as measured by certain standardized chemistry tests. He also concluded that the amount of chemistry retained was more closely related to achievement than to intelligence.

In the realm of pure research as opposed to technological research O'Kelley and Heyer have shown that a simple habit learned by male albino rats under a high degree of motivation is retained more effectively than a similar habit under a low degree of motivation.[18] A similar study involving human subjects showed similar significant results.[19]

Bruner, discussing human memory, states that "unless detail is placed into a structured pattern, it is rapidly forgotten".[20] He suggests that a vivid detail that carries the meaning of an event may

---

[16]McDougall, op. cit., p. 59.

[17]John R. Smeltz, "Retention of Learnings in High School Chemistry," The Science Teacher, 23, October 1956, p. 285.

[18]Lawrence I. O'Kelley and Albert W. Heyer, "Studies in Motivation and Retention," The Journal of Comparative and Physiological Psychology, 41, 1948, p. 466.

[19]Lawrence I O'Kelley and Albert W. Heyer, "Studies in Motivation and Retention: II. Retention of Nonsense Syllables Learned Under Different Degrees of Motivation," The Journal of Psychology, 27, 1949, p. 143.

[20]Jerome S. Bruner, The Process of Education (New York: Vintage Books, 1960), p. 24.

be a technique of condensation. "What learning general or fundamental principles does is to ensure that memory loss will not mean total loss, that what remains will permit us to reconstruct the details when needed."[21]

Bruner implies that a method involving discovery is a good means of instilling attitudes concerning the organization of student learning in such a way that learning is made useable and meaningful in thinking.[22] He implies that learning gained by a student through discovery may be more accessible to him in the future.

2.3 SUMMARY

This concludes the survey of the relevant literature. An attempt has been made to discuss methods comparison studies with respect to experimental improvements that might be incorporated within the design of the present study. This has drawn attention to the need for a more precise statement of educational objectives among other things.

Studies relevant to the retention of various educational outcomes were then cited. The findings of Tyler and McDougall provide support for the present experimental hypotheses concerning differential retention among educational objectives.

---

[21]Bruner, op. cit., p. 25.

[22]Bruner, op. cit., p. 20.

CHAPTER III

SPECIFICATION OF RELEVANT PREPARATIONS

3.0 INTRODUCTION

The test of the experimental hypotheses required the choice of an appropriate experimental design, the choice of relevant pretests, and the construction of pertinent unit tests. Prior to the inception of relevant procedures a number of decisions were to be made. This chapter concerns itself with these decisions and also with the essential remaining experimental preparations.

3.1 THE CONTENT OF THE UNIT

The hypotheses stated in Chapter I, section 1 indicated that an experimental unit from the revised grade nine science curriculum was to be used in the study. In order to reduce the possible Hawthorne effect it was felt that the unit should be one taught not at the beginning of the course but at some later date when the students had become accustomed to the teacher and also to the laboratory procedures. The heat unit appeared to be a suitable choice especially since the author who has to teach the unit was more familiar with physics than with the other scientific disciplines.

The unit on heat was prepared by the British Columbia Junior Secondary Science Revision Committee. There were 26 experiments listed in the unit with several of the experiments having a varying number of parts. The suggested time to be taken for the instruction of the unit was 22 class periods, each of one hour duration. Certain material in the Jr. Sec. Sc. Rev. Com. curriculum outline was deleted because it

did not appear to contribute materially to the understanding of the unit while in other instances further clarification of relevant concepts was sought by increasing the number of guide questions. Finally, a sequence of 17 lessons was developed, each lesson to be taught within a 55 minute class period.

The choice of material in the unit was not only governed by factors listed above but also by the equal treatment time factor. It was decided that both treatments (methods) were to have equal class time to minimize maturational and historical interference which could adversely affect the internal validity of the study. All of the basic experiments recommended by the revision committee and a number of the enrichment sections were included in the final unit. Several members of the Science Education department at the University of British Columbia, namely, Mrs. J. Woodrow, Mr. D. Webster, and Mr. H. Cannon, were asked to confirm the choice of material to be included in the unit.

The content of the unit was divided into eight sections which might be given the following titles:

1. The relationship between work and heat;
2. Sources of heat;
3. The expansion of solids, liquids, and gases;
4. Temperature and thermometers;
5. The measurement of heat;
6. The transference of heat energy;
7. Heat is a form of molecular motion; and
8. Heat causes a change of state.

The heat unit was taken from the Science 9 Experimental Edition of the 1966 curriculum bulletin.[1]

3.2 THE UNIT TESTS

a) Content of the Unit Tests. As indicated by the earlier discussion regarding objectives it was decided that the outcomes of interest were to be the Knowledge and Application categories of the Taxonomy. Since it has been generally accepted that the Taxonomy categories in the Handbook I: Cognitive Domain form a progression from simple to complex, and that each category contains the behaviors included in previous categories, it was felt that the choice of these objectives listed above was also desirable because of the degree of their categorical separation.

After thorough familiarization with the heat unit, the author designed the test items and classified them by objective. Several difficulties were noted during item classification; the first, that classification is relative to the nature of the instruction and the second, that classification is relative to the behavior considered critical in the item since most items involve more than one type of behavior. With respect to the former, Furst states:

> ..... that one must know or assume something about the nature of the students' prior experiences before one can classify a test item in a particular category. Thus, while on the surface a test item may seem to deal with the application of a principle, in the actual

---

[1]B. C. Department of Education, Junior Secondary School Science (Victoria, B. C. : The Queen's Printer, 1966).

[2]Edward J. Furst, Constructing Evaluation Instruments (New York: Longmans, Green and Co., 1958), p. 95.

> situation one could not be sure unless one knew whether the situation was in fact new to the students or whether it had been discussed previously. To consider the item as an application situation, one would have to assume that it was new to the students; otherwise it would fall in the recall-of-information category.[3]

To ensure the least number of errors in item classification several members of the Science Education department referred to earlier were asked to check the classifications.

Following this item classification two tests were to be constructed; one containing items evoking behaviors classifiable as Knowledge and the other containing items evoking behaviors classifiable as Application.

b) Type of Test. It was decided that each of the two tests should be power tests since the degree of speededness was of no interest as a factor in the study. Generous time limits were to be provided so that at least 90 percent of the students might attempt an answer to the last question on the test.

In comparison with other commonly used types of objective test items, multiple choice items are relatively high in ability to discriminate between better and poorer students. It was for this reason that multiple choice items were to be constructed for the two tests. The problem of the number of responses to each item was considered and it was felt that in the interests of analysis efficiency there should be an equal number of responses to each item. Further, five responses for each item would have been desirable from the standpoint of chance score reduction but four responses per item appeared a more practical objective especially in view of the construction difficulty of valid distracting responses.

---

[3] Ibid.

It was expected on a priori grounds that the mean difficulty index of the items on the Knowledge test would be greater than its counterpart on the Application test so that one might be required to allow a greater portion of time per question for the Application items. This would imply decreasing the number of useable Application items on the one test compared with the larger number of Knowledge items on the other. This disparity in item numbers between tests could almost certainly be expected to yield different reliability coefficients for the two tests, since the greater the number of items on an examination, other things being equal, the more reliable the scores obtained from it.

Professional test writers are able to produce achievement test reliability coefficients above .90 with fewer than one hundred items. However, it was deemed impractical to design eighty to one hundred items for each of the two tests in view of the short duration of "experimental" instruction and undesirable in view of the overlong testing time required.

The decision on the number of items to be included on each test, therefore was largely determined by the amount of time available for test administration within a class period. A prominent consideration in the decision was the balance attempted between increasing the number of items to increase reliability and decreasing the number of items to decrease speededness. A practical compromise was struck and it was decided that given a 50 minute test writing period, a test containing 40 items would be suitable on the basis of requirements mentioned previously.

3.3 THE PRETESTS

The concept of pretest use in experimental design has been well entrenched in the methodology of research workers in psychology and

education even though randomization between groups is recognized as the most adequate all-purpose assurance of lack of initial biases.[4] However, when the use of intact groups precludes randomization one may use statistical techniques to adjust group means on the dependent variable for variations among the groups on one or more concomitant variables provided that the latter correlate to some degree with the dependent variable. In order to provide for such equalization among intact groups to be used in this study, it was necessary to identify those factors which might be expected to affect the experimental dependent variable differentially.

It was felt that numerical ability, verbal reasoning, and prior general science knowledge were such factors. It is known that numerical ability and verbal reasoning are highly related to science achievement.[5] Evidence from the study of behaviors related to achievement would tend to suggest that many of the same behaviors may be responsible for retention.[6] The authors of the Taxonomy have noted that behaviors which can be generalized and applied in a number of different situations may be expected to exhibit greater permanence of learning than those which are so specific that they are likely to be encountered only rarely

---

[4]Donald T. Campbell and Julian C. Stanley, "Experimental and Quasi-Experimental Designs for Research on Teaching," Handbook of Research on Teaching, N. L. Gage, editor (Chicago: Rand McNally & Co., 1963), p. 195.

[5]G. K. Bennett, et. al., Differential Aptitude Tests Manual, (New York: The Psychological Corporation, 1963).

[6]Donald M. Medley and Harold E. Mitzel, "Measuring Classroom Behavior by Systematic Observation," Handbook of Research on Teaching, N. L. Gage, editor (Chicago: Rand McNally & Co., 1963), pp. 286 - 290.

throughout the educational program.[7]

Similarly, prior general science knowledge may be expected to influence retention on a priori grounds. If a student scores well on a measure of general science knowledge then his performance can be explained in terms of his manifestation of greater interest and motivation.[8]

Since standardized tests providing measures on the three criteria were readily available commercially it was decided that the following tests were to be administered as pretests prior to the instruction of the unit:

1. The Numerical Ability Test of the Differential Aptitude Tests;
2. The Verbal Reasoning Test of the Differential Aptitude Tests;[9]
3. The Read General Science Test: Evaluation and Adjustment Series.[10]

The Numerical Ability Test and the Verbal Reasoning Test were chosen for reasons cited above. The reliability coefficients range from .82 to .89 for very large samples of junior high school students throughout the United States. Both tests exhibit long term consistency to the degree exhibited by the following reliability coefficients. The Numerical Ability coefficients ranged from .74 to .75 whereas the Verbal Reasoning coefficients ranged from .82 to .87 for groups comparable to the ones

---

[7]Bloom, op. cit., p. 42.

[8]Pauline S. Sears and Ernest R. Hilgard, "The Teacher's Role in the Motivation of the Learner," Theories of Learning and Instruction, N.S.S.E. Yearbook, Part I (Chicago: The University of Chicago Press, 1964), p. 182.

[9]G. K. Bennett, et al., op. cit.

[10]John G. Read, Read General Science Test Manual (New York: World Book Co., 1951).

mentioned above.

One of the principal criticisms of the Differential Aptitude Test battery has been that intercorrelation between tests has not been as low as might be desirable. Ideally one would like to deal with pure instead of overlapping factors in order to account for a greater portion of variability in one's variable of interest. Carroll illustrates this hybrid nature of the tests when he suggests that the Verbal Reasoning Test probably measures a combination of the verbal ability and deductive reasoning factors.[11] The intercorrelation between the Numerical Ability Test and the Verbal Reasoning Test is reported to be .58 . Although this is higher than one might like it to be it is not expected to impair the validity of experimental measurement.

The Read General Science Test was chosen for its high physical science orientation. The test is reported to contain 42 percent physics items, 28 percent biology items, 4 percent chemistry items, and 26 percent general science items. The test was standardized on students whose median chronological age was that of ninth graders and the reliability coefficients on small samples reported by the test author ranged from .85 to .88 . In these respects it appeared to be the most suitable standardized test available.

Benjamin Bloom notes that the test attempts to measure knowledge of basic facts and principles of science as well as the ability to apply knowledge in problem solving situations.[12] With respect to these

---

[11]John B. Carroll, "Differential Aptitude Test Review", Fifth Mental Measurements Yearbook, Oscar K. Buros, editor (New Jersey: The Gryphon Press, 1959), p. 670 - 673.

[12]Benjamin S. Bloom, "Read General Science Test Review", Fourth Mental Measurements Yearbook, Oscar K. Buros, editor (New Jersey: The Gryphon Press, 1953), p. 628.

objectives it was felt that prior general science knowledge measured by the Read Test would correlate well with the measures obtained by the two unit tests and hence with the retention variable.

3.4 THE SUBJECTS

The subjects to be used in the investigation were members of two of the author's three grade nine classes at Britannia Secondary School in Vancouver, British Columbia. All of the students were on the academic technical program and most had attended the school during the previous year.

During the summer of 1966, prior to the timetabling of students to specific classes, efforts were made to assign students randomly to the five academic technical classes available. These efforts were unsuccessful as a result of subsequent timetabling difficulties. Although the use of intact groups was considered far from ideal practice it was nevertheless decided to proceed with intact groups since indirect or statistical control could remove potential sources of bias in the experiment.

The problem concerning which three of the five science classes were to be assigned to the author was essentially determined by timetabling considerations. Although randomization in assigning classes to teachers would have been preferred it is highly unlikely that any systematic bias would have been present in this assignment. It was therefore considered unlikely that the author's classes were not representative of the school's grade nine academic technical population.

There was no evidence available on optimum sample size so that an arbitrary limit of two classes appeared to provide a suitable number of

experimental subjects. At first the use of several teachers in the experiment was contemplated but this was abandoned for a variety of reasons. Among these were the apparent difficulties of achieving uniformity of teaching quality and uniformity of treatment. It was also felt that the addition of another dimension to the study without theoretical expectation of greater than average gains in efficiency design was an unsound scientific practice. Accordingly, it was decided to proceed with two groups of thirty-eight and thirty-nine students respectively, both groups being taught by the same teacher.

3.5 THE TRYOUT PROBLEM

Prior to the use of the final unit tests in the study, decisions had to be made concerning the number of items to be used, the difficulty and discrimination levels to be required and the reliability coefficients to be expected. To obtain information upon which to base these decisions tryout tests were to be planned. These tests were to identify weak or defective items and to provide the difficulty and discriminating power of individual items. The number of items on each tryout test could be expected to be considerably reduced when all decisions regarding the final tests were complete.

The time limit on the tryout tests was set at fifty minutes each, the same as that set for the final tests. To answer the items on the Knowledge test would require about one minute each. Thus a sample of 50 items was to constitute the Knowledge test and a somewhat smaller sample of 45 items was to constitute the Application test.

In accordance with information found in the literature it was at first thought that several schools should be involved in the

experimental tryout of the tests. Conrad suggests that the assessment of the adequacy of a tryout sample solely in terms of the number of students tested is inappropriate since school differences as well as pupil differences ought to be taken into account.[13]

A variety of reasons prevented the use of several schools in the tryout sample. A number of schools were either teaching parts of the grade nine revised science course or none at all. Others had either completed the instruction of the heat unit much earlier or were still in the process of completion. The former were not considered due to the well known effect of student indifference toward tests which are not credited toward final grades.

It was finally decided that a single school was to be used; one in which the grade nine students had just completed the study of the heat unit. It was felt that this procedure would certainly provide a better measure of the performance of the tests under conditions similar to those prescribed by this study's methodology than would procedures involving several schools.

Although the school was not randomly chosen there were no indications of systematic bias which might lead one to believe that any differences on relevant criteria between tryout and experimental samples were due to chance fluctuation. Further considerations concerning the tryout sample are given in Chapter 4, section 3.

---

[13]Herbert S. Conrad, "The Experimental Tryout of Test Materials," Educational Measurement, E. F. Lindquist, editor (Washington: American Council on Education, 1951), p. 253.

3.6 THE EXPERIMENTAL DESIGN

The experimental design chosen for use in this study was a variation of one which has frequently been used in educational research. The latter experimental design involves an experimental group and a control group both given a pretest and a post-test, but in which the two groups do not have pre-experimental sampling equivalence. The modification of this design concerns a shift in the time of treatment from between the test and retest to before the test, since the experimental variable of interest is retention.

When test-retest loss specific to the experimental or control group is to be explained by the experimental hypothesis, alternative hypotheses must be effectively eliminated. Such alternative hypotheses may arise from factors threatening the internal validity of a study. Factors listed by Campbell and Stanley such as the specific events occurring between the first and second measurement in addition to the experimental variable, the processes within the students operating as a function of time, the effects of taking a test upon the scores of a second testing, biases resulting in differential selection of students for the comparison groups, and differential selection due to experimental mortality were not seen as a threat to the internal validity of this study.[14] However, regression and interaction between specific selection differences distinguishing the two groups and the variables mentioned

---

[14]Donald T. Campbell and Julian C. Stanley, "Experimental and Quasi-Experimental Designs for Research on Teaching," Handbook of Research on Teaching, N. C. Gage, editor (Chicago: Rand McNally & Co., 1963), p. 175.

above could pose a serious threat to internal validity. It was decided to control the potential difficulty of regression by analysis of covariance and it was felt that interaction would not present difficulty in the present study due to the occurrence of both tests following the treatment.

According to Campbell and Stanley the factors which may jeopardize external validity or generalizability are the interaction effect of testing, the interaction effects of selection biases and the experimental variable, and the reactive effects of experimental arrangements.[15] The first of these in which a pretest might increase or decrease the student's responsiveness to the experimental variable would not be considered of great importance in this study because the pretest has followed the treatment. In order for this factor to pose a threat the likelihood of a student becoming more highly motivated to engage in additional learning by the test rather than by the instruction which preceded it should be reasonably high. Naturally, a basic requirement for the above argument is the lack of awareness on the part of the student that he will write the same test at a later date.

The second factor concerns the possibility that the experimental effects which may validly be demonstrated hold only for the unique population from which the two groups were jointly selected. Thus, although one's results may be internally valid the difficulty lies in the attempts to generalize the results to the population of interest. There is no convenient way to resolve this problem although it has been suggested that the requirements of the present design place fewer limita-

---

[15] Ibid.

tions on sampling than do similar designs thus partially reducing the potential threat to external validity.[16]

The last factor concerns unrepresentativeness brought about by artificiality of experimental setting. A student's knowledge that he is participating in an experiment is expected to have an extraneous effect on his behavior.

In the present study one of the objectives was the removal of any artificiality in the teaching methods through deliberate teacher consideration of and provision for previously observed classroom circumstances as outlined in Chapter 4, section 1, so that a comparison of two methods actually found in practice might be made without compromise toward requirements of experimental design. Further, it was felt that conditioning of the experimental group prior to the commencement of the experiment should minimize the novelty of the experimental setting.

This concludes the account of the preparations essential to procedural implementation and allows the description of experimental methodology.

---

[16] Ibid., p. 220.

CHAPTER IV

METHODOLOGY

4.0 INTRODUCTION

This chapter deals with the preparation of the tryout and final tests as well as the administration and results of the pretests, the tryout tests and the final tests. The methodology pertinent to the preparation of the students and the instruction of the unit is also discussed.

4.1 PREPARATION OF THE STUDENTS

It was decided that it was necessary to familiarize the prospective members of the experimental groups with the procedures followed in scientific investigation. The familiarization procedures did not appear to favor subsequent learning by any particular group.

Each of the author's three grade nine classes was divided into small groups at the beginning of the year and about four or five periods were spent in small group discussion. Various aspects of scientific method, deductive reasoning and inference were discussed by the students and reports made. All persons in the three groups were given the same familiarization treatment.

The students were taught a unit on chemistry in which they were frequently called upon to use parts of the information gathered during the group discussions. Experimentation was performed on a rotation basis with several small groups engaged in discovery under supervision at any one time. The students not called upon to perform an experiment on a particular day solved chemistry problems at their desks.

By late October the three classes were randomly assigned to three teaching methods in a preliminary investigation of the effects of instructional methods on achievement. The three methods were a discovery method, a demonstration method, and a lecture method. The unit taught was one on forces, taken from the revised grade nine science course. The members of the discovery and demonstration groups received instruction sheets related to the experimental laboratory work daily. The primary difference between these two methods used at this time was the demonstration of the experiment by the teacher for one class and not for the other. Otherwise conditions were similar in that both answered the same questions and formulated their own experimental conclusions.

An achievement test of 40 items constructed to measure a variety of objectives ranging from Knowledge to Analysis was administered to all three groups following the completion of the unit.[1] The results were rather inconclusive since the group test means found were in close proximity. The demonstration group test mean was slightly larger than either of the remaining group means. No statistical analysis of the means was carried out.

Following the instruction of the unit on forces all classes were again taught by a combination of methods, that is, lecture, discussion, and demonstration methods. The discovery method was not used however. The unit taught in this way dealt with magnetic and electrical properties of matter.

---

[1]Bloom, op. cit.

4.2 THE INSTRUCTION OF THE UNIT

The class taught by the lecture method during the preliminary investigation was randomly deleted and hence was not used in the experiment on retention. The group formerly taught by the self discovery method was now taught using the same method whereas the class taught by the demonstration method was now taught using a modified version of the demonstration method.

Part of the instruction of the unit rests upon the completion of a sequence of experiences. Some notion of the presentation used may be obtained by examining the content of the sample lessons reproduced in appendix B.

Any equipment new to the students in the experimental group was demonstrated but its use in the solution of the problem was not discussed. Equipment was then distributed and the students proceeded to perform the experiment. The teacher meanwhile was free to circulate among the groups of students. If difficulties arose, the students were asked questions to bring out ideas which might be useful in the solution or clarification of the difficulty.

Approximately ten to fifteen minutes before the end of each period the students were reminded to begin formulating conclusions if they had not already begun. The last three minutes of the period were relegated to clean-up activities. The students submitted laboratory reports at the end of each period.

The laboratory reports were checked daily and helpful comments written where required. These reports were then returned to the students at the beginning of the following period. It was thought that a daily check would maximize motivational factors and also lead to more efficient

use of student time.

Reports were also collected from the control group whenever a demonstration was performed and graded in a manner similar to that used for the experimental group.

A number of topics received only lecture treatment in the lecture demonstration treatment while others received both lecture and demonstration treatment. The lectures were generally of ten to fifteen minutes in length and were usually followed by student note making. The points raised in the lecture occasionally served as starting points for class discussion. Other points were illustrated by means of diagramatic analysis at the blackboard.

Members of the author's thesis committee acted as judges of the teaching procedures used in order to confirm the implementation of the two methods as defined in Chapter I.

4.3 PREPARATION OF THE TRYOUT TESTS

It was evident that no commercial achievement tests on the heat unit were available for the purposes of this study. This necessitated the preparation of tryout tests and subsequently, the final tests.

First, the Taxonomy was used in making an outline of objectives defining the ways in which the students were to deal with various parts of unit content. Following this a table of specifications was prepared on the basis of the statement of objectives and the outline of unit content. This table was then used in the preparation of the two sets of items. An abbreviated form of this table may be seen in Table III.

In the preparation of the items appropriate evaluation situations were sought so that individual items would reflect the attainment or

non-attainment of the relevant educational objective. The wording of the items was carefully scrutinized to prevent the inclusion of unintentional clues. Item distractors were constructed on the basis of the author's experience of student errors. This was considered more effective in some cases than in others since it was thought that discussions with students and corrected laboratory reports revealed only a fraction of the misconceptions held by students. Frequent consultation of Ebel's Measuring Educational Achievement provided numerous helpful suggestions for writing the multiple choice items.[2] The Taxonomy and Gerberich's Specimen Objective Test Items were also consulted in order to minimize classification errors.[3]

The number of items within a content section of each test was roughly proportional to the amount of class time devoted to the instruction of the content section. Thus the two tests were similar in item proportions within content sections.

A total number of 97 items was constructed; 52 belonging to the Knowledge category and 45 to the Application category. Several members of the Science Education department then checked the items for content validity, scientific accuracy, and correctness of classification. Two items from each category were deleted for reasons of classification disagreement and unforeseen item weaknesses leaving 50 Knowledge items and 43 Application items.

The items were then classified with respect to difficulty by

---

[2]Robert L. Ebel, Measuring Educational Achievement (Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1965), pp. 151 - 200.

[3]J. R. Gerberich, Specimen Objective Test Items (New York: Longmans Green & Co., 1957).

the author. A broad classification of three groups ranging from easy to difficult was first established for the items of each test. A finer classification containing five groups was then established and the items within each group were re-examined to determine their likely position on this crude scale.

Items within each test were subsequently matched on content and difficulty, each member of the matched pair being randomly assigned to an odd or even position on the tryout test. This was done to facilitate the calculation of reliability coefficients. The items of each test were arranged in order of increasing difficulty; a practice which is commonly followed in test construction. These arrangements were examined to determine the presence of possible correct response patterns and to detect the presence of unequal correct response proportions among the four alternatives. Changes in item position were then effected to remedy the imperfections noted.

4.4 ADMINISTRATION OF THE TRYOUT TESTS

a) The problem of guessing. Upon considering whether to correct the scores on the tryout tests for guessing it was noted that most experimental studies on the subject have shown that the effect of announced correction for guessing has been a very slight improvement on the reliability and validity of the scores.[4] Thus, corrected scores will rank students in approximately the same relative positions as uncorrected scores. Also, it is not considered bad pedagogy to accept

---

[4]Ebel, op. cit., p. 227.

rational guessing since one frequently finds himself in a situation requiring a decision but without sufficient evidence upon which to base that decision. Further, the correction formula assumption that all wrong answers are the result of blind guessing is untenable since numerous exceptions may be noted. Also, on a speeded test one could expect slower students to guess blindly on the items near the end of the test but since both tryout tests are power tests a correction for guessing would be much less useful than for a speeded test.

All things considered, then, it was decided that no correction for guessing should be made but that a well worded set of instructions should be adopted to encourage making optimum use of partial information but to advise against blind guessing.[5]

The same test directions were used for the tryout tests as for the final tests. These directions are reproduced in appendix A. Test instructions concerning time limits were given to the tryout test administrators verbally. It was later reported by the test administrators that all instructions had been carried out.

b) The results of the tryout tests. The two tryout tests were written by the members of six grade nine classes from Montgomery Junior Secondary School in Coquitlam. A total number of 160 students wrote the tests.

With few exceptions, most students attempted all of the items on the Knowledge test. There was no evidence to indicate that the test

---

[5]Instructions adapted from T. D. M. McKie, "An Investigation of the Relationship between the Relevance Category of Achievement Test Items and their Indices of Discrimination," (unpublished Master's Thesis, The University of British Columbia, 1962), pp. 81 - 82.

had not functioned as a power test. However, there was evidence of speededness on the Application test. Eight percent of the students did not complete the last seven items. Admittedly, these had been judged the most difficult items but more than ninety five percent of the students had attempted all earlier items.

Also, there was evidence that despite the instructions, considerable guessing took place. In this case the guessing problem is more likely to be associated with the Application test than with the Knowledge test.

For each test, the papers were scored and the students' performance on each item recorded, including omissions. The papers were then ranked in order of decreasing score and divided into three groups to facilitate the calculation of the Johnson discrimination index.[6] This procedure requires the separation of the top 27 percent and the bottom 27 percent of the ranked test papers from the middle 46 percent.

Difficulty indices were computed for each item and are shown in Table I. The difficulty index of an item is the proportion of students in the sample who answered the item correctly.

Discrimination indices were calculated for each item and are shown in Table II. The discrimination index used in this study is the difference in proportion of correct responses between the group scoring in the top 27 percent on the total test and the group scoring in the bottom 27 percent on the same test.

---

[6]A. Pemberton Johnson, "Notes on a Suggested Index of Item Validity: The U - L Index," Journal of Educational Psychology, 42 (1951), pp. 499 - 504.

TABLE I

DISTRIBUTION OF DIFFICULTY INDICES[1] COMPUTED FROM TRYOUT TESTS

Difficulty Indices (percents)	Number of Items	
	Knowledge	Application
80 - 84	2	
75 - 79	3	1
70 - 74	1	1
65 - 69	4	1
60 - 64	5	1
55 - 59	11	1
50 - 54	4	4
45 - 49	6	2
40 - 44	2	2
35 - 39	3	8
30 - 34	3	6
25 - 29	1	5
20 - 24	3	6
15 - 19	1	3
10 - 14	1	2
N =	50	43
Mean Difficulty Index (percent) =	51.8	41.0

[1]Corrections have been made for omissions.

TABLE II

DISTRIBUTION OF DISCRIMINATION INDICES[1] COMPUTED FROM TRYOUT TESTS

Discrimination Indices	Number of Items	
	Knowledge	Application
.70 - .74	1	
.65 - .69		
.60 - .64	2	
.55 - .59	2	
.50 - .54	3	1
.45 - .49	5	3
.40 - .44	5	5
.35 - .39	12	3
.30 - .34	5	7
.25 - .29	5	10
.20 - .24	1	4
.15 - .19	6	3
.10 - .14		2
.05 - .09	1	3
.00 - .04		
N =	48 [2]	41 [2]
Mean Discrimination Index ==	.354	.274

[1]Corrections have not been made for omissions.

[2]Two items from each test discriminated negatively and thus were not included in this table.

4.5 THE ADMINISTRATION OF THE PRETESTS

The students were informed well beforehand that the three tests (to be used as covariates) were to be written. They were urged to do as well as they could and were assured that test performance did not dictate failure in the course.

Each of the pretests was administered on a separate day during regular class periods. Timetable difficulties prevented the two classes involved in the experiment from writing the tests at the same time and since the classes were not scheduled into successive blocks it was necessary to maximize test security arrangements by administering the test first to one of the groups during their regular science period and then to the second during a period in which they were to receive instruction in another subject. The author was able to administer the tests himself. It was assumed that there was little or no information passed between classes since the time between classes was a short four minute period and since the students recognized that test performance did not dictate success or failure in the course.

4.6 THE PREPARATION AND ADMINISTRATION OF THE FINAL TEST

a) The selection of the items. The difficulty index distributions of the two tryout tests were markedly different. Similarity between these two distributions would have facilitated greatly the testing of the third experimental hypothesis. Since this was not the case an alternative procedure to be described in section 5.5 was used.

It is known that a test will provide a maximum number of discriminations among test candidates if the test items are uncorrelated and if they are all of 50 percent difficulty. Further, if the items

are all perfectly correlated, the number of discriminations made by all of the items will be identical with the number made by one item of 50 percent difficulty. Clearly, when student rank on some criterion is desired, it is to one's advantage to design a test in which item difficulty indices cluster as closely as possible around the 50 percent level since item inter-correlations are generally rather low.

However, the point has been made that not only the theoretical soundness of this procedure should be considered in the construction of achievement tests but also its psychological soundness. With reference to the latter one need only consider the plight of the duller than average youngster who writes the test, proceeding item by item, only to sense that his chance of failure on each item is greater than his chance of success. It is not difficult to see how this might lead to more guessing and lower test reliability coefficients.

A possible solution was to distribute items about a somewhat higher value than 50. (60. was aimed at), and to avoid items of extremely low difficulty indices.

Two items on the Knowledge test discriminated negatively, that is to say they were discriminating in a direction opposite to that of the remaining items of the test, and were discarded. Knowledge items with difficulty indices below 34.0 were then discarded except for one item whose difficulty index was 19.0 and whose discrimination index was sufficiently high to warrant its inclusion. This left a total of forty items with a mean difficulty index of 53.9 .

Two items on the Application test also had negative discrimination indices and were discarded. Nine items with difficulty indices below 25.0 were then discarded leaving only one item with a difficulty index

below 25.0 . Thirty two items remained at this point but the difficulty index distribution was rather positively skewed, that is, the distribution was not symmetric, and had a disproportionate number of difficult items. Thus the choice lay in rejecting even more items with almost certain loss of potential reliability or in writing additional items to provide a more symmetric distribution. Since there was insufficient time to prepare and administer a second tryout test, ten new items were written and were administered to three of the six classes involved in the tryout sample. Six of the original thirty two items were then replaced with six of the new items, thus providing a reasonably symmetric difficulty index distribution with a mean of 46.2 .

The discrepancy between the means of the difficulty index distributions for the two tests indicated that the two sets of test items could not be regarded as samples drawn from populations with the same distribution and could therefore not be used directly to test the hypothesis of differential retention. It was therefore necessary to choose matched items from the two tests on the basis of the performance by the students in the experimental and control groups. This procedure will be described in Chapter V.

b) The construction of the unit test. It was reported that in each tryout test, items had been matched on the basis of content and expected difficulty level. Each of the final tests was processed in this manner using the items selected from the tryout tests. However, on this occasion matching could be accomplished on the basis of difficulty indices obtained from the tryout sample and this, it was expected, would increase test reliability to a considerable extent. The members of the paired items were again randomly assigned to an odd

or even position on the test.

Table III indicates the item distribution by content category in the two unit tests.

The items on each test were then arranged in order of increasing difficulty, checked for equal distribution of correct responses, and the alternative choices dispersed in random fashion to prevent correct response patterns. The tests are reproduced in appendix C.

The time limits for both tests were set at fifty minutes. Upon consideration of the tryout test results and the number of questions on each test it was believed that all students would have time to consider all items thus making the tests true power tests.

c) The administration of the unit tests. The students in the experimental and control groups were familiar with the procedures of responding to multiple choice tests since they had written several of these tests during the year. The strategy of test writing had been discussed with the students throughout the year so that it was felt that blind guessing could be curbed to some considerable degree.

The Knowledge test was administered to both groups during the same period on the morning of the day following the completion of the heat unit. Similarly, the Application test was administered in the same manner on the school day following the writing of the first test. All students indicated that they were able to consider all items on both tests. It appeared, judging from the omitted items, that the instructions concerning guessing had been observed.

Two students were absent during the writing of the Knowledge test and four during the writing of the Application test.

TABLE III

ITEM DISTRIBUTION BY CONTENT CATEGORY
IN THE TWO UNIT TESTS

CONTENT CATEGORY	Number of Items		Percentage of Total	
	Knowledge	Application	Knowledge	Application
Work and Energy Changes	2	2	5.	6.2
Heat Sources	3	0	7.5	0
Expansion	5	7	12.5	22.
Temperature and Thermometers	4	3	10.	9.3
Measurement of Heat	6	5	15.	15.5
Heat Transfer	8	7	20.	22.
Molecular Motion	4	4	10.	12.5
Change of State	8	4	20.	12.5
TOTAL	40	32		

4.7 THE ADMINISTRATION OF THE RETEST

a) Administration of the Retest. The students had been informed several times throughout the year that they would be examined on various parts of the course toward the latter part of the school year. This was done in an attempt to curb possible hostility toward surprise retesting which was to occur approximately six to seven weeks following the first unit test administration. The students were also told that these tests given late in the school year would be counted toward their final grade in science. However, attention was not specifically drawn to the retesting of the heat unit.

Six weeks following the first test administration the students were informed during their morning roll call period that they were to write a science test during the first two class periods of that morning. It had been decided to administer both unit tests at the same time since the variable of interest, namely retention, would in all likelihood be greatly affected by any student foreknowledge of retesting.

The students were then randomly assigned to the writing of either the Knowledge or the Application test during the first period. This was done to confound potential practice or order effects. During the second period the students who had written the Knowledge test now wrote the Application test and vice versa. The time limit of fifty minutes for each test was again strictly observed.

b) The test results. There were five absentees on the day the retests were written. Altogether, data on the Knowledge criterion was obtained for thirty three students in the experimental group and thirty eight students in the control group. Data on the Application criterion

was obtained for thirty two students in the experimental group and thirty seven students in the control group.

It appeared from the number of items left out that the instructions regarding guessing had been observed.

The test scores were processed by the University of British Columbia Computing Centre.

CHAPTER V

THE STATISTICAL ANALYSIS

5.0 INTRODUCTION

In this chapter the unit test scores, the pretest scores, and the retest scores are presented together with test reliabilities. Two tests of significance are made and interpreted in the light of the assumptions of the analysis of variance and covariance model. The selection of items constituting the subtests which are used to test the differential retention hypothesis is described and a test of significance is made.

5.1 THE UNIT TEST RESULTS AND RELIABILITIES

a) The unit test scores. The distribution of scores for the experimental and control groups on the two unit tests may be seen in Table IV.

It can readily be seen from Table IV that the means and medians for the control group are greater than the corresponding means and medians for the experimental group. It cannot be asserted that these differences are due to chance since the groups were not initially chosen randomly. Tests of the retention hypotheses will need to take account of such initial differences. The statistical "equalization" of groups with respect to the testing of experimental hypotheses will be discussed later in this chapter.

The difficulty and discrimination indices for the two unit tests are shown in Table V and the loss scores for the two groups on each test are shown in Table VI. The latter will be used in testing the retention hypothesis.

TABLE IV

DISTRIBUTION OF UNIT TEST SCORES

Test Scores	Frequency			
	Knowledge		Application	
	Exp. Group	Cont. Group	Exp. Group	Cont. Group
36 - 37	1	3		
34 - 35	1	1		
32 - 33	6	5		
30 - 31	2	9		
28 - 29	3	3		
26 - 27	8	6		2
24 - 25	6		2	1
22 - 23	2	1	2	6
20 - 21	2	3	2	1
18 - 19	4	3	4	9
16 - 17		4	8	6
14 - 15	2		6	6
12 - 13			6	2
10 - 11			3	3
8 - 9			2	2
N =	37	38	35	38
M =	25.9	27.2	15.8	17.4
Mdn =	26.1	28.8	15.6	17.5
S^2 =	33.80		21.22	
S =	5.81		4.61	

TABLE V

DISTRIBUTION OF DIFFICULTY AND DISCRIMINATION INDICES FOR THE UNIT TESTS

Difficulty Indices	Frequency		Discrimination Indices	Frequency	
	Knowl.	App.		Knowl.	App.
.93 - .97	3		.63 - .67		2
.88 - .92	3	1	.58 - .62	3	
.83 - .87	2	1	.53 - .57	3	1
.78 - .82	4		.48 - .52	5	3
.73 - .77	4	2	.43 - .47	2	3
.68 - .72	3	5	.38 - .42	6	6
.63 - .67	4	2	.33 - .37	5	3
.58 - .62	6	1	.28 - .32	3	4
.53 - .57	5	4	.23 - .27	4	2
.48 - .52	2		.18 - .22	3	4
.43 - .47	1	5	.13 - .17	4	2
.38 - .42	1	7	.08 - .12	2	
.33 - .37	1	2	.03 - .07		
.28 - .32	1		(-).02 - .02		1
.23 - .27			(-).07 - (-).03		1
.18 - .22		2			
N =	40	32	N =	40	32
M =	.68	.53	M =	.35	.34

TABLE VI

DISTRIBUTION OF LOSS SCORES

Loss Scores	Frequency			
	Knowledge		Application	
	Exp. Group	Cont. Group	Exp. Group	Cont. Group
-7		1	1	
-6	1		1	
-5	1	1	7	2
-4	3	3	1	2
-3	1	3	3	6
-2	5	3	2	5
-1	2	7	3	4
0	7	6	3	3
1	4	5	5	4
2	2	3	2	2
3	2		2	4
4	1	3	1	
5	1			2
6	1	1	1	1
7	2	1		1
8		1		1
N =	33	38	32	37
M =	.15	-.05	-1.34	.03
S^2 =	5.40	5.56	4.44	5.56
S =	2.32	2.36	2.11	2.36

b) The reliability of the unit tests. The construction of the unit tests has already been described in Chapter IV. The items within each test were matched on the basis of content and difficulty. The test reliability of each test was determined by correlating the scores on the two halves and subsequently using the results in the Spearman-Brown formula[1]

$$r_{11} = \frac{2r_{\frac{11}{22}}}{1 + r_{\frac{11}{22}}}$$

The reliability coefficients (r_{11}) obtained for the Knowledge and Application tests were .82 and .80 respectively. They were estimated by the Spearman-Brown formula from the half - length coefficients ($r_{\frac{11}{22}}$) which were .69 and .67 respectively.

The test reliability coefficients were not as high as desired but were accepted as reasonably good in view of the small sample size. As a matter of interest, the Application test reliability coefficient, calculated by the Spearman-Brown formula, becomes .83 if the length of the test is increased to forty items, as in the Knowledge test. Of course, it is assumed that the eight additional items exhibit the same characteristics as do the thirty two items of the actual test.

c) The correlation between the tests. The correlation coefficient between the Knowledge and Application tests was .72 . This provided a measure of the degree of relationship between the two sets of test scores. It was inferred that in so doing, the correlation coefficient

---

[1]Robert L. Ebel, op. cit., p. 315.

showed the extent to which the two tests measured different aspects of mental functioning. This inference was made rather cautiously since it is so highly dependent upon test validity. This correlation is rather higher than most of those reported in the literature[2] though not as high as some[3].

An assumed true correlation coefficient r = .72 yields the following partitioning of Application test score variance:

(i) error variance : 20 percent;

(ii) variance explainable by the variability of Knowledge test scores : 52 percent; and

(iii) true variance, reliably measured and different from the variability of Knowledge test scores : 28 percent.

Evidently the two kinds of performances involve some common abilities, but each requires certain abilities that the other does not. It is of course assumed that the test items have all been accurately classified according to the specifications of the Taxonomy.

5.2 THE PRETEST RESULTS

The distributions of scores on the three pretests may be seen in Tables VII and VIII.

The control group means were greater than the experimental group

---

[2]Edward J. Furst, "Effect of the Organization of Learning Experiences upon the Organization of Learning Outcomes," Journal of Experimental Education, 18 (1954).

[3]T. R. McConnell, "A Study of the Extent of Measurement of Differential Objectives of Instruction," in An Appraisal of Techniques of Education: Symposium (Washington: American Educational Research Association, National Education Association, February, 1940).

TABLE VII

DISTRIBUTION OF PREPEST SCORES

Test Scores	Frequency			
	Verbal Reasoning Test		Numerical Ability Test	
	Exp. Group	Cont. Group	Exp. Group	Cont. Group
45 - 46		1		
43 - 44	1			
41 - 42	1	4		
39 - 40	1	3		1
37 - 38		3	2	8
35 - 36	2	3	5	3
33 - 34	5	1	1	9
31 - 32	6	7	4	3
29 - 30	1	3	4	5
27 - 28	2	3	5	4
25 - 26	2	3	2	1
23 - 24	3	1	6	1
21 - 22	3	3	1	1
19 - 20		2	1	
17 - 18	1		1	1
15 - 16			1	1
13 - 14	3			
11 - 12	2			
9 - 10		1		
N =	33	38	33	38
M =	27.3	31.1	28.2	31.7

TABLE VIII

DISTRIBUTION OF PRETEST SCORES

Read General Science Test Scores	Frequency	
	Exp. Group	Cont. Group
69 - 70		1
67 - 68		1
65 - 66		
63 - 64	1	
61 - 62		1
59 - 60	1	
57 - 58	1	2
55 - 56		4
53 - 54	1	4
51 - 52		6
49 - 50	3	1
47 - 48	4	1
45 - 46	3	5
43 - 44		2
41 - 42	6	
39 - 40	3	5
37 - 38	5	1
35 - 36	2	2
33 - 34		1
31 - 32		1
29 - 30		
27 - 28	1	
25 - 26		
23 - 24		
21 - 22		
19 - 20		
17 - 18	2	
N =	33	38
M =	42.2	48.4

means for each of the three pretests. It was therefore evident that the two groups could not be regarded as random samples from a normally distributed population and that indirect or statistical control would need to be exercised in the "equalization" of the groups prior to comparison.

5.3 THE STATISTICAL TEST OF HYPOTHESIS I

The statistical hypothesis to be tested was the null hypothesis, $H_0 : \mu_E' = \mu_C'$, the alternative hypothesis being $H_1 : \mu_E' < \mu_C'$, where μ' represents an adjusted mean in the analysis of covariance sense. Since $t^2_{(\nu)} = F_{(1, \nu)}$, where 1 and ν are the respective degrees of freedom of the numerator and denominator in the F ratio obtained in the analysis of variance, a one-tailed $t = \sqrt{F}$ test was to be made. The level of significance for the F ratio was preset at 5 percent.

Loss scores, the difference between initial and final unit test scores, were calculated for each student on the basis of the Knowledge test results. The experimental and control groups' scores were plotted separately against their original Knowledge test scores, their Numerical Ability scores, their Verbal Reasoning scores, and their Read General Science Test scores. The scatter plot graphs indicated small positive correlations between Knowledge loss scores and both Numerical Ability scores and original Knowledge scores. Knowledge loss scores and Read General Science scores indicated a small negative correlation. However, Knowledge loss scores were correlated positively with Verbal Reasoning scores for the control group and negatively for the experimental group. This impaired the value of the Verbal Reasoning factor as a covariate

and led to its rejection in the testing of the first hypothesis.

The remaining scatter plots were examined for linearity of regression and there was no reason to believe that any of these relationships was nonlinear. Initial knowledge of the unit as measured by the Knowledge test, numerical ability as measured by the Numerical Ability test, and prior general science knowledge as measured by the Read General Science test were then chosen as covariates to be used in the testing of the first hypothesis.

There was no reason to believe that within class regression coefficients differed to a degree greater than that expected by chance. Both skewness and variance disparity were in evidence to a small degree among the experimental and control group Application loss scores but neither was in evidence among the Knowledge loss scores. It was felt that possible violation of the normality and homogeneity of residual variance assumptions would pose no threat to hypothesis test validity since the F test in the analysis of covariance is known to be robust with respect to these two violations.[4]

In the analysis of covariance procedure group means are first adjusted on the basis of performance on the covariates. This adjustment is necessary to "equalize" the groups and eliminate any selection bias which may have been in operation.

The analysis of covariance was performed and produced the following results. The adjusted means for the control group and the experimental group were .04 and .05 respectively.

---

[4]B. J. Winer, Statistical Principles in Experimental Design (New York: McGraw Hill Book Co., 1962), p. 586.

Source	df	SS (Adj.)	MS (Adj.)	F
Groups	1	.002	.002	0.00
Error	66	704.07	10.67	

Clearly, since F is less than 1 , the null hypothesis cannot be rejected.

In view of this result, it was concluded that there was no statistical basis for maintaining that there is greater mean retention at the knowledge level among students taught by means of a discovery method than among those taught by means of a lecture demonstration method.

All loss scores and covariate scores for the two groups are shown in appendix D.

5.4 THE STATISTICAL TEST OF HYPOTHESIS II

The statistical hypothesis to be tested was the null hypothesis, $H_0 : \mu_E' = \mu_C'$, the alternative hypothesis being $H_1 : \mu_E' < \mu_C'$, where μ' represents an adjusted mean. A one-tailed $t = \sqrt{F}$ test was to be made and the level of significance for the F ratio was to be set at 5 percent.

Loss scores were calculated for each student on the basis of the Application test results. The experimental and control groups' loss scores were again plotted separately against the original Application test scores, the Numerical Ability test scores, the Verbal Reasoning test scores, and the Read General Science test scores. The results indicated a positive correlation between Application loss scores and

original Application scores, a negative correlation between Application loss scores and Numerical Ability scores and a small negative correlation between Application loss scores and Read General Science test scores. There appeared to be no correlation between Application loss scores and Verbal Reasoning test scores. Therefore, the covariates used in the test of the first hypothesis were also used in the test of the second.

The remaining data was then examined as described in section 5.3 to determine whether or not the analysis of covariance assumptions had been met. These appeared to be satisfied and so the analysis was performed. The adjusted means obtained for the control group and the experimental group were .34 and -1.70 respectively. The results of the analysis of covariance are shown in the following table.

Source	df	SS (Adj.)	MS (Adj.)	F
Group	1	61.58	61.58	6.79 **
Error	64	580.37	9.07	

** $p < .01$ for a one-tailed $\underline{t}$ test.

Since $t_{(64)} = \sqrt{F_{(1,\ 64)}} = 2.61$ and the tabled value of $t_{(64)}$ at the 1 percent level is 2.390 the difference in means is not only significant at the 5 percent level but also well beyond the 1 percent level. Thus the null hypothesis was rejected and the alternative hypothesis, $H_1 : \mu'_E < \mu'_C$, was accepted.

It was therefore concluded that there was a statistical basis for maintaining that there is greater mean retention at the application

level among students taught by means of a discovery method than among those taught by means of a lecture demonstration method.

The obtained value of t is significant at the 1 percent level so that that there is little chance of this result being a type I error, though this is always a possibility which only replication will resolve.

5.5 THE LOSS SCORE RELIABILITIES

The reliability of the difference between two scores can be expressed in a simple formula, which reads

$$r_{diff} = \frac{\frac{r_{11} + r_{22}}{2} - r_{12}}{1 - r_{12}}$$

where r_{11} is the reliability of one measure,

r_{22} is the reliability of the other measure, and

r_{12} is the correlation between the two measures.[5]

The reliability coefficient for the Knowledge loss scores obtained by using the Knowledge test reliability coefficient, .82, and the pre-post Knowledge test correlation, also .82, in the above formula is 0.0 . The reliability coefficient for the Application loss scores, obtained in the same manner using the Application test reliability coefficient, .80, and the pre-post Application test correlation, .73, is .26 .

Much larger loss score reliability coefficients had been hoped for. This reduction of reliability is of course a tremendous challenge

---

[5]Robert L. Thorndike and Elizabeth Hagen, Measurement and Evaluation in Psychology and Education (New York: John Wiley & Sons, Inc., 1961), p. 192.

in studies involving loss or gain scores. The fact that such clear-cut results were obtained in the case of the Application hypothesis is quite remarkable in view of the severe measurement error, and renders the conclusion of more than ordinary interest.

5.6 THE SUBTESTS USED IN THE TEST OF HYPOTHESIS III

In order to test the hypothesis concerning differential retention it was necessary to provide comparable tests, matched with respect to content and difficulty. The low reliabilities of the two tryout tests prevented their use in the construction of the tests to be used in a test of this hypothesis. Instead, items were chosen from the two final unit tests on the basis of difficulty index and discrimination index comparability.

The distribution of the eighteen items thus selected is shown in Table IX. Most of the matched item difficulty indices were within .07 of each other with the greatest difference being .13 . The mean difficulty indices of the Knowledge and Application subtests were .63 and .62 respectively; the mean discrimination indices, .39 and .34 respectively. The difficulty index distributions, shown in Table X, had the same range, approximately the same form, reasonable symmetry, and approximately equal variability.

The reliability coefficients for the Knowledge and Application subtests, obtained by the procedure mentioned in section 5.1, were .67 and .57 respectively. These were low in comparison to normal test reliability figures, but were considered reasonable in view of the small number of items used.

TABLE IX

DISTRIBUTION OF ITEMS BY CONTENT CATEGORY IN THE SUBTESTS

Content Category	Number of Items	
	Knowledge Subtest	Application Subtest
1	4	3
2		1
3	2	3
4	1	1
5	2	1
6	3	2
7	2	2
8	4	5

MATCHED SUBTEST ITEMS

Item No. on Knowledge Test	Item No. on Application Test
3	19
5	1
8	14
9	3
11	4
16	10
18	26
19	27
22	20
24	12
25	16
28	5
29	9
31	15
33	28
34	29
36	8
37	11

TABLE X

DISTRIBUTION OF DIFFICULTY INDICES
IN THE SUBTESTS

Difficulty Indices	Frequency	
	Knowledge Subtest	Application Subtest
.88 - .92	1	1
.83 - .87	1	1
.78 - .82	1	
.73 - .77	3	2
.68 - .72	1	2
.63 - .67	1	4
.58 - .62	4	1
.53 - .57	1	3
.48 - .52	2	
.43 - .47	1	
.38 - .42		3
.33 - .37	2	1
N =	18	18
M =	.63	.62
S^2 =	.033	.025
S =	.18	.16

5.7 THE STATISTICAL TEST OF HYPOTHESIS III

The statistical hypothesis to be tested was the null hypothesis, $H_0 : \mu_E = \mu_C$, the alternative hypothesis being $H_1 : \mu_E < \mu_C$, where μ is an unadjusted mean. This required a one-tailed test. The level of significance had again been set at 5 percent.

Loss scores were calculated on both the Knowledge subtest and the Application subtest for each student. Next, mean loss scores were calculated for both experimental and control groups on each of the two subtests. Each of the four mean loss scores was divided by its own standard error of difference and the four terms were combined to form experimental and control group mean differential loss scores. Finally, each of the two mean differential loss scores was divided by its standard error of difference. A test was then to be made to determine whether the difference in the respective scaled mean differential loss scores of the experimental and control groups was significant.

The division of each difference by its standard error accomplished the transformation of all loss scores to a similar scale, that is, all differences were scaled to their respective standard deviations. This scaling procedure allowed the formation of the differences between Application and Knowledge mean loss scores to be used in testing the differential retention hypothesis.

The derivation and interpretation of the above terms and the test statistic are given in appendix E.[6]

No adjustment to "equalize" groups was possible using this

---

[6]Derivation based on personal communication with Dr. T. D. M. McKie, The University of British Columbia.

statistical procedure. Lack of adjustment was not considered a serious problem in this case since initial bias appeared to favor the control group and hence would not prejudice possible significant results in favor of the experimental group.

The test statistic,

$$Z = \frac{\left\{\dfrac{\bar{A}_{E_2} - \bar{A}_{E_1}}{\sigma_{A_{E_2} - A_{E_1}}} - \dfrac{\bar{K}_{E_2} - \bar{K}_{E_1}}{\sigma_{K_{E_2} - K_{E_1}}}\right\} - \left\{\dfrac{\bar{A}_{C_2} - \bar{A}_{C_1}}{\sigma_{A_{C_2} - A_{C_1}}} - \dfrac{\bar{K}_{C_2} - \bar{K}_{C_1}}{\sigma_{K_{C_2} - K_{C_1}}}\right\}}{\sqrt{2\left\{\dfrac{1 - \rho_E}{N_E} + \dfrac{1 - \rho_C}{N_C}\right\}}}$$

$$= \frac{\left\{\dfrac{-\frac{22}{32}}{2.08} - \dfrac{-\frac{3}{32}}{1.98}\right\} - \left\{\dfrac{\frac{6}{37}}{2.72} - \dfrac{\frac{6}{37}}{2.14}\right\}}{\sqrt{2\left\{\dfrac{1 - .091}{32} + \dfrac{1 - (-.176)}{37}\right\}}},$$

was calculated and its value found to be -.768 , the negative sign referring to a net gain instead of loss. For the one-tailed test, the probability of obtaining a value of $\underline{Z}$ less than -.768 was determined from the table of the normal probability integral to be .2212 .

The null hypothesis could not be rejected and was therefore accepted. It was concluded that there was no statistical basis for maintaining that there was greater mean differential retention among students taught by means of a discovery method than among those taught by means of a lecture demonstration method.

CHAPTER VI

CONCLUSIONS AND SUMMARY

6.0 INTRODUCTION

This study was designed to examine the effects of discovery learning on retention. This was achieved by comparing a discovery method of instruction with a demonstration method on the basis of retention of objectives defined by the Taxonomy. There were, however, several limitations of the study which will be discussed prior to the statement of conclusions.

6.1 LIMITATIONS OF THE STUDY

First, there are inferential difficulties in that only a cautious generalization of the experimental results to the population of interest may be made. If one attempts to generalize to the population of British Columbia high school science students one must consider the limitations resulting from the use of a single content unit, a single teacher, and a small sample. However, in so far as the students in the sample are representative of the population ---- and there is no evidence of nonrepresentativeness ---- the results may be generalized to this population. The requirements of a strengthened generalization demand greater sample size and more representative sampling.

Secondly, the lack of randomization necessitated the use of covariates in the equation of experimental and control groups. Since the correlations between the covariates and the dependent variable were rather small there was some question as to whether all of the pre-experimental bias in favor of the control group was removed. Although

this effect operated to strengthen the results obtained, it should be noted that in the event of replication of this study the existence of possible covariates more highly correlated with loss scores than the present covariates ought to be investigated.

As implied in the previous paragraph, the low reliability of the loss scores was undoubtedly responsible for the concealment of a sizeable portion of the experimental effect. Improvement of loss score reliability might have been achieved through increasing the number of test items. Decreasing the correlation between test and retest would also help, but unfortunately it is not within the control of the experimenter.

6.2 CONCLUSIONS

First, there is no statistical basis for asserting that discovery learning results in greater retention at the Knowledge level of mental functioning than does learning acquired by means of a lecture demonstration method. In view of the loss score reliability on the Knowledge criterion these results are to be expected.

Secondly, there is a statistical basis for asserting that discovery learning results in greater retention at the Application level than does learning acquired by means of a lecture demonstration method. The statistical test shows the difference between the experimental and control groups to be significant beyond the 1 percent level although the level set for null hypothesis rejection was only 5 percent. An effect of this magnitude is all the more remarkable when considered in the light of the low reliabilities obtained for the Application loss scores.

Behaviors involving Application are probably more frequently encountered during discovery learning than during learning obtained by

means of a lecture demonstration method. This would appear to give a reasonable explanation of the results obtained, namely, the absence of a significant difference on the Knowledge retention criterion and the presence of a significant difference on the Application retention criterion.

Finally, there is no statistical basis for the belief that discovery learning results in greater differential retention than does learning acquired by means of a lecture demonstration method. The results obtained are in the right direction, however, and this effect would be enhanced if a correction could have been made to remove the initial biases of the groups. Conceivably the effect might have been large enough to be not attributable to sampling fluctuation.

6.3 IMPLICATIONS OF THE STUDY

Despite the lack of loss score reliability, clear cut results were obtained for the Application hypothesis. This finding supports the earlier Tyler studies which claimed that Application behaviors were more readily retained than Knowledge behaviors and provides justification for the current movement in science education concerning course revision. Clearly, if discovery learning contributes to the development of greater retention among Application objectives than among Knowledge objectives, then a greater emphasis on Application objectives is warranted. Further, if Application behaviors are more likely to be encountered than Knowledge behaviors during discovery learning, then science courses designed to be taught using discovery methods hold a decided advantage over the old science courses.

6.4 POSSIBILITIES FOR FURTHER RESEARCH

There are several possibilities for further research. The first concerns a replication of the present study to investigate more thoroughly differential retention between Knowledge and Application objectives. A greater number of items -- perhaps about 60 -- matched with respect to content and difficulty should be constructed and reliability coefficients in excess of .90 should be aimed at for both tests. This problem, because of its magnitude, might well be undertaken by an institutional research group rather than by a single individual. Further investigation of the scaling problems involved in making such comparisons is needed, and perhaps also of the associated covariance adjustment problem.

Investigation of differential retention between other objectives may also be carried out.

Finally, a great deal of research might profitably be directed toward the systematic design of instructional methods based on learning principles.

6.5 SUMMARY OF THE INVESTIGATION

It was hypothesized on a priori grounds that:

1. greater mean retention would occur at the Knowledge and Application levels among students taught by a discovery method than among students taught by a lecture demonstration method; and
2. greater mean retention would occur at the Application level than at the Knowledge level among students taught by a discovery method than among students taught by a lecture

demonstration method.

Each of two ninth grade science classes in a single school was taught a heat unit using one of the methods mentioned above. The instructional methods were assigned randomly to intact classes both handled by the same teacher.

Two multiple choice achievement tests covering the content of the heat unit were constructed; one consisting of items in the Knowledge category and the other consisting of items in the Application category. A tryout of each of these tests was conducted upon 160 students in a single school thus allowing the elimination of unsuitable items. The resulting unit tests, with forty Knowledge and thirty two Application items respectively, were administered to the students of the experimental and control groups both immediately following and six weeks following the conclusion of the heat unit.

The reliability coefficients of the Knowledge and Application tests, estimated by correlating the half test scores and applying the Spearman-Brown formula, were .82 and .80 respectively.

Possible concomitant variables such as prior general science knowledge, verbal reasoning, and numerical ability were measured by appropriate pretests. Covariates were chosen on the basis of pretest score correlation with loss scores. The analysis of covariance was performed and the results provided a statistical basis for the acceptance of the experimental hypothesis dealing with the retention of application objectives and for the rejection of the experimental hypothesis dealing with the retention of knowledge objectives.

Items were matched on the basis of content and difficulty index to form Knowledge and Application subtests. A normal Z statistic

was used to test the differential retention hypothesis and the results indicated that there was no statistical basis for the acceptance of this experimental hypothesis.

BIBLIOGRAPHY

A. BOOKS

Bloom, Benjamin S., et al. Taxonomy of Educational Objectives, Handbook I: Cognitive Domain. New York: Longmans, Green and Co., 1956.

Bloom, Benjamin S. "Read General Science Test Review," Fourth Mental Measurements Yearbook, Oscar K. Buros, editor. New Jersey: The Gryphon Press, 1953.

Bruner, Jerome S. The Process of Education. New York: Vintage Books, 1960.

Campbell, Donald T., and Julian C. Stanley. "Experimental and Quasi-Experimental Designs for Research on Teaching," Handbook of Research on Teaching, N. L. Gage, editor. Chicago: Rand McNally and Co., 1963.

Carroll, John B. "Differential Aptitude Test Review," Fifth Mental Measurements Yearbook, Oscar K. Buros, editor. New Jersey: The Gryphon Press, 1959.

Conrad, Herbert S. "The Experimental Tryout of Test Materials," Educational Measurement, E. F. Lindquist, editor. Washington: Amer. Council on Educ., 1951.

Ebel, Robert L. Measuring Educational Achievement. New Jersey: Prentice Hall, Inc., 1965.

Edwards, Allan L. Experimental Design in Psychological Research. New York: Holt, Rinehart and Winston, 1965.

Furst, Edward J. Constructing Evaluation Instruments. New York: Longmans, Green and Co., 1958.

Gerberich, J. R. Specimen Objective Test Items. New York: Longmans, Green and Co., 1957.

Medley, Donald M., and Harold E. Mitzel. "Measuring Classroom Behavior by Systematic Observation," Handbook of Research on Teaching, N. L. Gage, editor. Chicago: Rand McNally and Co., 1963.

Sears, Pauline S., and Ernest R. Hilgard. "The Teacher's Role in the Motivation of the Learner," Theories of Learning and Instruction, N. S. S. E. Yearbook, Part I, E. R. Hilgard, editor. Chicago: The University of Chicago Press, 1964.

Thorndike, Robert L., and Elizabeth Hagen. Measurement and Evaluation in Psychology and Education. New York: John Wiley and Sons, 1961.

Tyler, Ralph W. "The Relation between Recall and Higher Mental Processes," Education as Cultivation of the Higher Mental Processes, C. H. Judd, editor. New York: The MacMillan Co., 1936.

Wallen, Norman E., and Robert M. W. Travers. "Analysis and Investigation of Teaching Methods," Handbook of Research on Teaching, N. L. Gage, editor. Chicago: Rand McNally and Co., 1963.

Watson, Fletcher G. "Research on Science Teaching," Handbook of Research on Teaching, N. L. Gage, editor. Chicago: Rand McNally and Co., 1963.

Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw Hill Book Co., 1962.

B. PUBLICATIONS OF GOVERNMENT AND OTHER ORGANIZATIONS

Bennett, G. K., et al. Differential Aptitude Tests Manual. New York: The Psychological Corp., 1963.

British Columbia Department of Education. Science Curriculum Bulletin. Victoria, B. C.: The Queen's Printer, 1963.

British Columbia Department of Education. Science Curriculum Bulletin. Victoria, B. C.: The Queen's Printer, 1965.

British Columbia Department of Education. Junior Secondary School Science. Victoria, B. C.: The Queen's Printer, 1966.

McConnell, T. R. "A Study of the Extent of Measurement of Differential Objectives of Instruction," in An Appraisal of Techniques of Education: Symposium. Washington: Amer. Educ. Research Assoc., N. E. A., 1940.

Read, John G. Read General Science Test Manual. New York: World Book Co., 1951.

U. S. Department of Health, Education, and Welfare, Office of Education, New Dimensions in Education (Effectiveness in Teaching No. 2). Washington: 1960,

C. PERIODICALS

Cunningham, H. A. "Lecture method versus individual laboratory method in science teaching - A Summary," Science Education, 30 (1946), 70-82.

Furst, Edward J. "Effect of the Organization of Learning Experiences Upon the Organization of Learning Outcomes," Journal of Experimental Education, 18 (1950), 215.

Johnson, A. Pemberton. "Notes on a Suggested Index of Item Validity: The U-L Index," Journal of Educational Psychology, 42 (1951), 499-504.

McDougall, William P. "Differential Retention of Course Outcomes in Educational Psychology," Journal of Educational Psychology, 49, No. 2, (April, 1958), 53.

O'Kelley, L. I., and A. W. Heyer. "Studies in Motivation and Retention," The Journal of Comparative and Physiological Psychology, 41 (1948), 466.

O'Kelley, L. I., and A. W. Heyer. Studies in Motivation and Retention: II. Retention of Nonsense Syllables Learned under Different Degrees of Motivation," The Journal of Psychology, 27 (1949), 143.

Smeltz, John R. "Retention of Learnings in High School Chemistry," The Science Teacher, 23 (October, 1956), 285.

Suchman, Richard J. "Inquiry Training: Building Skills for Autonomous Discovery," Merrill-Palmer Quarterly, 7 (July, 1961), 147.

Tyler, Ralph W. "Permanence of Learning," Journal of Higher Education, IV (April, 1933) 203-204.

D. UNPUBLISHED MATERIALS

Boldt, Walter B. "Grade Placement of an Experimental Unit in Secondary School Physics." Unpublished Master's thesis, The University of British Columbia, Vancouver, 1963.

McKie, T. D. M. "An Investigation of the Relationship between the Relevance Category of Achievement Test Items and their Indices of Discrimination." Unpublished Master's thesis, The University of British Columbia, Vancouver, 1962.

APPENDIX A

DIRECTIONS TO STUDENTS

1. Many of the items on this test will appear different from any that you have seen before. Most of them require a little bit of thinking as well as just remembering something that you have learned. However, ALL of them CAN be done by any student who has learned the work done to date WELL. So however strange or hard a question may appear at first sight, remember --- YOU KNOW ENOUGH TO ANSWER IT; all it needs is your knowledge plus a little thought!

2. You may answer questions even when you are not completely sure that your answers are correct. In such cases, intelligent consideration of the choices provided may help you to gain marks. HOWEVER, you should AVOID WILD GUESSING as this may result in a reduction in your score.

3. Give each question careful thought, but work as quickly as you can. If you find a question too difficult, do not linger over-long on it. Pass on to the next ones, and return later to any that you have missed.

4. For each question there are four possible answers. You are to decide which is the best one. Mark only ONE of the four choices, as shown in the sample item below.

   ITEM: The chemical symbol for mercury is:

   a) M b) Mg c) Hg d) Ag

   The correct answer is c), so make a heavy mark in the space marked c). Erase completely any answer you may wish to change, and be careful not to make any stray marks of any kind on your answer sheet or on your test booklet.

   a) b) c) d)

5. DO NOT OPEN THE TEST BOOKLET UNTIL YOU ARE TOLD TO START! The working time for this test is 50 minutes.

6. As soon as you receive your answer sheet, complete the required information on the top (i.e. NAME, SCHOOL, etc.).

7. On tests of this kind, it is not expected that any student will get all of the items correct; in fact, it is usual for most average students to score around 50%. The pass-mark will undoubtedly be much less than 50%, so if you have to leave quite a few questions unanswered, do not worry -- and DO NOT GUESS.

APPENDIX B

SAMPLE LESSON FOR THE DISCOVERY METHOD

A brief review of the previous lesson is conducted.

The aim of this lesson is to study the effect of external pressure on the boiling point of water.

The teacher asks questions designed to probe student understanding of the definition of the terms and the purpose of the lesson. For example, "What is meant by the terms 'external pressure' and 'boiling point'?" and "Describe in your own words the problem to be investigated."

The teacher asks questions concerning the formulation of possible hypotheses and the testing of these hypotheses. For example, "If the external pressure is increased, would you expect the boiling point to increase, decrease, or stay the same?" and "How could the hypothesis be tested?"

Further student questioning elicits more specific information concerning hypothesis testing. For example, "If you expect the boiling point to increase when the external pressure increases, how are you going to arrange your equipment and what procedures will you carry out to check your hypothesis?"

Instead of following up this questioning by explanation at this point the teacher allows the students to decide for themselves which hypothesis to test and discuss among themselves the procedures to be followed and the results anticipated.

The teacher checks that the students can handle the apparatus and then sets them to work while he circulates among the groups to provide assistance where necessary.

The students perform the experiment and record their observations. Students may have questions such as the following: "If the water

temperature is lower than 100° C. why does the water boil?" The teacher responds by asking a series of questions designed to lead the student to an understanding of and solution to his problem. For example, "What substance occupies the space above the water?" , "How is this substance affected when it loses heat?" , and "Is it as difficult for water molecules to move into the space above the liquid water after cooling the substance as it is before cooling it?"

The students receive some degree of assistance in their investigation from the guide questions on the experiment instruction sheets.

After testing one hypothesis a student may formulate and test alternative hypotheses. For example, instead of investigating the effects of decreasing the pressure above a liquid he may investigate the effects of increasing the pressure above the liquid.

The students formulate their conclusions, submit their reports, and return their laboratory equipment.

SAMPLE LESSON FOR THE LECTURE DEMONSTRATION METHOD

A brief review of the previous lesson is conducted.

The aim of this lesson is to study the effect of external pressure on the boiling point of water.

The teacher outlines by a lecture and blackboard presentation the theory involved in the lesson. For example, the students are asked to consider the effects of increasing and decreasing the pressure above a liquid. Molecular behavior is discussed with respect to both situations but the heating or cooling of the gas above the liquid is not mentioned. The students record the blackboard presentation in their notebooks.

The teacher asks questions concerning the definitions of terms and the purpose of the lesson as in the "discovery method". He also asks questions concerning the formulation of possible hypotheses and the testing of these hypotheses. The students discuss with the teacher which of the hypotheses relate directly to the problem to be investigated and what information is to be gathered to test these hypotheses.

The teacher directs the discussion toward the hypothesis which might most suitably be tested under existing conditions. In this case, for example, investigating the effects of decreasing the pressure above a liquid would be most suitable for purposes of demonstration.

Following the discussion, the teacher outlines the object, the hypothesis, and the plan of the investigation for the students on the blackboard. The students record this outline as part of their laboratory report.

The teacher performs the experiment and the students record their own observations. Wherever possible an experimental report format has

been provided for the students. This involves the systematic tabulation of data.

The students are required to give written answers to a series of questions as part of the laboratory report. These questions pertain to noteworthy points in the demonstration. For example, "When the flask is cooled is the pressure of the water vapor above the liquid equal to atmospheric pressure?"

The students formulate their own conclusions and submit their report.

APPENDIX C

KNOWLEDGE TEST

ON HEAT UNIT

1. The specific heat of an object is the amount of heat one gram of the object

   a) absorbs without changing its temperature.
   b) radiates before its temperature can be raised by one degree C.
   c) requires to raise its temperature by one degree C.
   d) contains after its temperature is raised by one degree C.

2. When a liquid is heated, the molecules

   a) move closer together.
   b) move into a definite pattern.
   c) become heat energy.
   d) move faster.

3. The temperature at which water changes from liquid to solid (when the barometric pressure is 76 cm. of mercury) is

   a) 212° F.
   b) 100° F.
   c) 32° F.
   d) 0° F.

4. When a warm piece of metal is placed in cold water, the heat lost by the metal is equal to

   a) the heat lost by the water.
   b) the heat absorbed by the water.
   c) the heat absorbed by the metal.
   d) the work done in changing the temperature of the water.

5. When matter absorbs radiation, the radiation is converted to

   a) heat.
   b) chemical energy.
   c) work.
   d) temperature.

6. Water and alcohol tend to

   a) cool at the same rate.
   b) expand at different rates.
   c) contract at the same rate.
   d) cool at the same rate and expand at the same rate.

7. Which of the following conduct heat?

   a) solids and gases only.
   b) liquids and gases only.
   c) solids and liquids only.
   d) gases, solids, and liquids.

8. Which of the following substances will have its temperature raised by one Centigrade degree when one calorie of heat is added to one gram of the substance?

a) water
b) turpentine
c) glycerol
d) alcohol

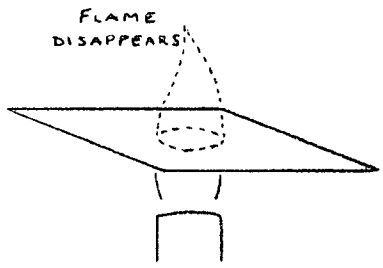
9. Which of the following statements is correct?

a) Water does not conduct heat.
b) Water is a poor conductor of heat.
c) Water is a moderate conductor of heat.
d) Water is a good conductor of heat.

10. Which of the following statements is correct?

a) Mercury does not conduct heat.
b) Mercury and water conduct heat equally well.
c) Mercury is a poorer heat conductor than water.
d) Mercury is a better heat conductor than water.

11. A copper gauze, positioned as shown, causes a Bunsen flame to disappear because



a) the flame cannot get through the gauze.
b) the heat is conducted away too rapidly by the copper gauze.
c) the natural gas has already been used up.
d) the heat is radiated away too rapidly by the flame.

12. Which of the following are sources of heat?
1. Sunlight 2. Burning fuels 3. Electric sparks

a) 1 only
b) 2 only
c) 1 and 2 only
d) 1, 2, and 3

13. The number of degrees between the freezing point and the boiling point of water on a Fahrenheit thermometer is

a) 0
b) 100
c) 180
d) 212

14. Which of the following is an example of a good conductor of heat?

a) zinc
b) water
c) china
d) pyrex

15. A flask containing water is fitted with a stopper and glass tubing as shown. When the flask is heated, the water level in the tube drops slightly because

a) water contracts when heated.
b) water expands when heated.
c) the flask expands faster than the water.
d) the flask contracts slower than the water.

16. When heat is added to a cubic foot of air, its volume

a) remains the same and its temperature increases.
b) increases and its temperature increases.
c) decreases and its temperature increases.
d) increases and its temperature remains the same.

17. When water freezes it

a) gives out heat.
b) takes in heat.
c) neither gives out nor takes in heat.
d) may either give out or take in heat.

18. When the temperature is above the freezing point of water, the numerical reading on the Fahrenheit thermometer

a) is proportional to the reading on the Celsius thermometer.
b) is greater than the reading on the Celsius thermometer.
c) is less than the reading on the Celsius thermometer.
d) may be greater than or less than the reading on the Celsius thermometer.

19. A stone falling from a height of six feet above the ground possesses a maximum of kinetic energy

a) at a height of six feet above the ground.
b) at a height of three feet above the ground.
c) just before it strikes the ground.
d) just after it strikes the ground.

20. The amount of heat required to raise the temperature of 1 gram of water from 99° C. to 100° C.

a) is greater than the amount of heat required to raise the temperature of 1 gram of water from 0° C. to 1° C.
b) is less than the amount of heat required to raise the temperature of 1 gram of water from 0° C. to 1° C.
c) is the same as the amount of heat required to raise the temperature of 1 gram of water from 0° C. to 1° C.
d) is the same as the amount of heat required to raise the temperature of 1 gram of ice from 0° C. to 1° C.

21. Bodies which are warmer than their surroundings are called

a) heat sources
b) heat indicators
c) radiators
d) conductors

22. For which of the following method(s) of heat transfer is no material medium required?

a) conduction
b) convection
c) radiation
d) conduction and radiation

23. Two aluminum blocks, one twice as heavy as the other, are both at the same temperature. The ratio of the amount of heat contained by the heavier block to the amount of heat contained by the lighter block is

a) 1:2
b) 1:1
c) 2:1
d) 4:1

24. Water cascading over a dam illustrates the following changes in forms of energy

a) Mechanical to kinetic to electrical
b) Potential to kinetic to heat
c) Kinetic to mechanical to heat
d) Kinetic to heat to potential

25. The operation of a compound bar thermostat is based upon the fact that

a) Metals expand when heated.
b) Different metals have different heat capacities.
c) Metals bend more easily when heated.
d) Different metals expand at different rates.

26. A thick walled glass jar and a thin walled pyrex beaker are heated by a Bunsen burner. Which of the following statements is most correct?

a) The jar breaks because it does not expand uniformly.
b) The beaker does not break because it does not expand uniformly.
c) The jar breaks because it expands uniformly.
d) The jar breaks because its walls are thicker.

27. When an impure substance dissolves in a liquid

a) the boiling temperature increases.
b) the freezing temperature increases.
c) the boiling temperature decreases.
d) the freezing temperature is not changed.

28. Decreasing the vapor pressure above water decreases its

a) boiling temperature.
b) water pressure.
c) boiling pressure.
d) water temperature.

29. The amount of heat an object possesses depends on the object's

a) temperature only.
b) mass and weight.
c) weight and temperature.
d) temperature and mass.

30. The specific heat of water is greater than

a) alcohol.
b) turpentine.
c) both alcohol and turpentine.
d) neither alcohol nor turpentine.

31. Heat applied to water at 100 C. tends to

a) raise the temperature of the water.
b) prevent heat loss from the boiling water.
c) raise the temperature of the steam.
d) change the state of the water.

32. Which of the following statements is correct?

a) Water boils at 90° C.
b) Water boils at 100° C.
c) Water boils at 110° C.
d) Water may boil at 90° C., or 100° C., or 110° C.

33. When water changes from a liquid to a gas

a) heat is given out.
b) heat is taken in.
c) heat is neither given out nor taken in.
d) heat is either given out or taken in.

34. The boiling point of water is not affected by

a) adding a salt which does not dissolve in water.
b) increasing the vapor pressure above the water.
c) adding table salt to the water.
d) decreasing the vapor pressure above the water.

35. Winds transfer heat by

a) conduction.
b) convection.
c) radiation.
d) convection and conduction.

36. The energy of molecular motion appears as

a) friction.
b) heat.
c) temperature.
d) potential energy.

37. Heat energy is radiated through space by

a) molecular collisions.
b) molecular group transfer.
c) molecular motion.
d) a process which does not depend on molecular motion.

38. An outdoor thermometer uses mercury for which of the following reasons?
    1. Mercury has a high boiling point.
    2. Mercury's freezing point is below that of water.
    3. Mercury expands as the temperature increases.
    4. Mercury does not stick to glass.

    a) 2 and 3 only
    b) 2, 3, and 4 only
    c) 1, 2, and 3 only
    d) 1, 2, 3, and 4

39. Which of the following statements is correct? Heat is usually produced when

    a) work is done.
    b) potential energy is entirely converted to kinetic energy.
    c) kinetic energy is entirely converted to potential energy.
    d) potential energy is lost.

40. Which of the following statements is correct?

    a) Shiny surfaces absorb energy better than black surfaces.
    b) Black surfaces reflect energy better than shiny surfaces.
    c) Black surfaces radiate and absorb energy better than shiny surfaces.
    d) Shiny surfaces radiate energy better than black surfaces.

APPLICATION TEST

ON HEAT UNIT

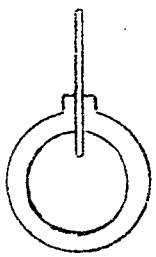
1. Wet clothing on a washline may require a long time to dry because

   a) the surrounding air molecules are not in motion.
   b) the specific heat of air is too low.
   c) the air pressure is too high.
   d) the amount of water vapor in the air is too high.

2. Cork is a poor conductor of heat because

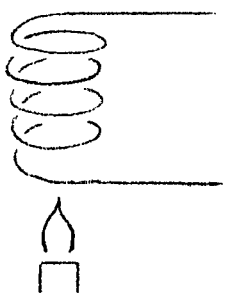
   a) it contains many air spaces.
   b) its molecules are separated by large distances,
   c) cork molecules do not absorb heat readily.
   d) it is a good reflector of heat.

3. A wire is sealed into an evacuated bulb made of pyrex glass. The bulb, which is to be dipped into liquid air at a very low temperature (-200 ° C.) is also to be used at room temperature (+20 ° C.). The most important factor to consider when designing this apparatus is



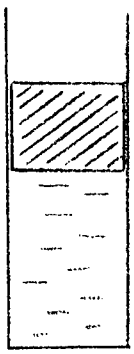
   a) the rate of thermal expansion of the wire and the glass.
   b) the thickness of the glass.
   c) the shape of the bulb.
   d) the thickness of the wire.

4. Water enters the coil



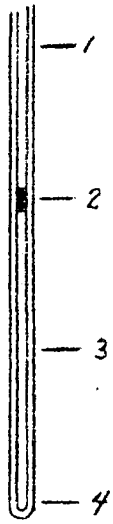
   a) at the bottom and rises as it cools.
   b) at the bottom and rises as it warms up.
   c) at the top and sinks as it cools.
   d) at the top and sinks as it warms up.

5. A cork, whose diameter is slightly smaller than that of the tube, is allowed to slide freely in a tube partially filled with water. If the water begins to boil the cork will be forced up by



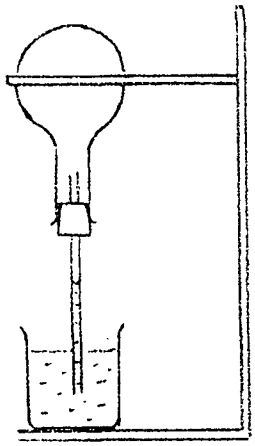
   a) air pressure.
   b) water pressure.
   c) atmospheric pressure.
   d) vapor pressure.

6. A small bubble of mercury is introduced into a piece of glass tubing one end of which is sealed. After the sealed end is placed in boiling water the bubble is at the position shown. After the sealed end is placed in a mixture of ice and water the bubble will be at



a) position 1.
b) position 2.
c) position 3.
d) position 4.

7. A flask filled with air at room temperature is inverted in a beaker of water as shown. When the flask is heated the level of the water in the beaker will



a) rise.
b) fall.
c) neither rise nor fall.
d) first rise, then fall.

8. A short distance from the spout of a steaming kettle the air contains a mist of water droplets, but the air between this mist and the spout is clear. The reason for this is

a) that water vapor is always invisible.
b) that all of the water molecules in this region are in the gaseous state.
c) that there is no form of water present in this region.
d) that the droplets in this region are travelling too fast to be seen.

9. Under normal conditions liquid water boils at 100 ° C. and liquid nitrogen boils at -196 ° C. When equal amounts of water (at 99 ° C.) and <u>liquid</u> nitrogen (at -197 ° C.) are poured together

a) both the liquid nitrogen and water freeze.
b) the liquid nitrogen freezes.
c) the water freezes.
d) neither the liquid nitrogen nor the water freeze.

10. A certain metal ring cannot be slipped onto a glass rod because the diameter of the glass rod is a little too large. It is possible to put the ring on the glass rod by

a) heating the rod.
b) cooling the ring.
c) cooling the rod.
d) cooling both the ring and the rod.

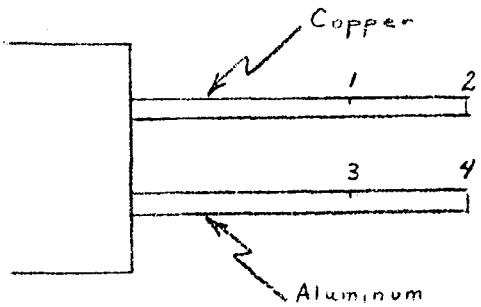
11. Electricity passing through the coil of an electric kettle causes water in the kettle to boil because electrical energy is converted to

a) chemical energy.
b) kinetic energy.
c) work.
d) potential energy.

12. A wooden block is sent sliding across a table top. The kinetic energy of the block is converted to

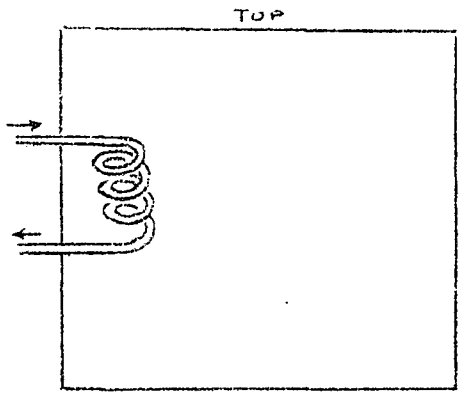
a) heat energy.
b) friction.
c) potential energy.
d) friction and potential energy.

13. Two rods, one made of aluminum and the other of copper, are initially attached to a plate held at a constant temperature of 80° C. If both rods are initially at 20° C. then



a) points 2 and 4 receive the same amount of heat at the same time.
b) points 2 and 3 receive the same amount of heat at the same time.
c) points 1 and 4 receive the same amount of heat at the same time.
d) points 1 and 3 do not receive the same amount of heat at the same time.

14. Water at 0° C. is circulated through a coil which is surrounded by water at 80° C. The temperature of the water in the tank

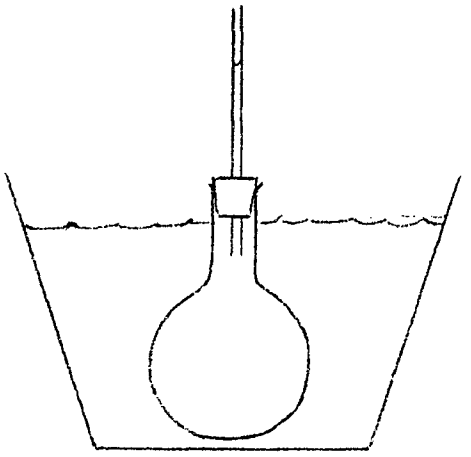


a) rises because work is done pumping water through the coil.
b) falls because heat is absorbed by the coil.
c) remains the same because the heat lost is equal to the heat gained.
d) remains the same because heat is radiated by the tank.

15. One hundred grams of ice at 0° C. are added to one hundred grams of water at 0° C. in an insulated container which allows no heat losses. The mixture will

a) become ice at 0° C.
b) become water at 0° C.
c) become mostly ice at 0° C.
d) remain the same.

16. A flask containing water is fitted with a stopper and glass tubing. The flask is placed in a hot water bath and the level in the tubing rises to a height of 10 inches. A mixture of 50 percent water and and 50 percent alcohol is poured into the same flask and the flask is again placed into the hot water bath. The level of the liquid in the tubing rises to a height of



a) less than 10 inches because different liquids are used.
b) 10 inches because the same amount of heat is absorbed.
c) more than 10 inches because different liquids are used.
d) 10 inches because the hot water bath is at a constant temperature.

17. A bar is made of a strip of iron and a strip of brass bonded together. For a one degree rise in temperature brass expands $1\frac{1}{2}$ times as much as iron. If one end of the bar is clamped as shown and the temperature is lowered fifty degrees the bar will

a) remain straight because iron is stronger than brass.
b) curve downward.
c) remain straight because the temperature is lowered.
d) curve upward.

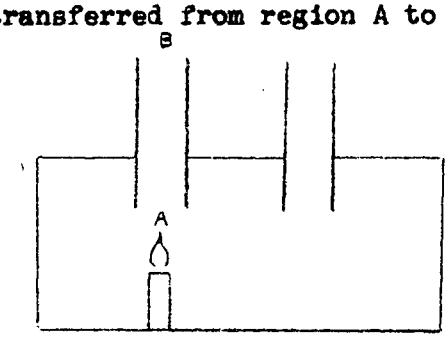
18. One end, X, of a closed tube filled with air is briefly cooled with an ice pack. When the ice pack is removed, the temperature will

a) rise rapidly at X.
b) rise rapidly at Y.
c) fall slowly at X.
d) fall slowly at Y.

19. A reading of 45 degrees on the Celsius (Centigrade) scale is the same as a Fahrenheit reading of

a) 57 degrees.
b) 77 degrees.
c) 81 degrees.
d) 113 degrees.

20. A closed box has two chimneys inserted at the top. A burning candle is placed underneath one of the chimneys as shown. Region A is just above the flame and region B is just above the chimney. Heat is transferred from region A to region B



a) mainly by convection and conduction.
b) only by conduction and radiation.
c) mainly by convection and radiation.
d) only by conduction and convection.

21. A 500 cm. rod changes 0.1 cm. in length for a one degree change in temperature (Celsius). If the rod's original temperature of 250 ° C. changes to 200 ° C. the length of the rod is

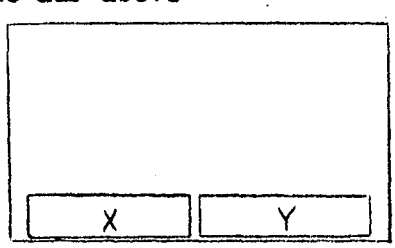
a) 450.0 cm.
b) 495.0 cm.
c) 499.5 cm.
d) 505.0 cm.

22. Water at 0 C. is circulated through a coil which is surrounded by water at 80 C. Which of the following statements is correct?

See diagram for #14.

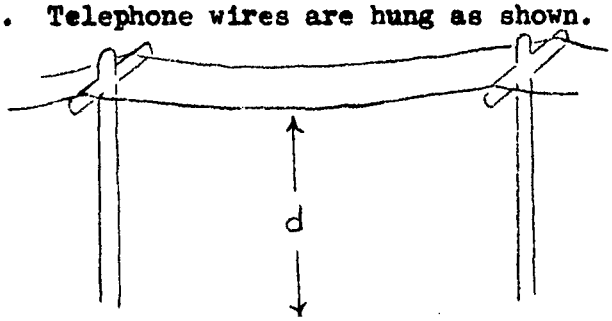
a) Water below the coil neither rises nor falls.
b) Water below the coil rises.
c) Water above the coil rises.
d) Water below the coil falls.

23. Materials X and Y, both at 80 C., are placed in a container filled with air at 20 C. If material X contains more heat than material Y the air above



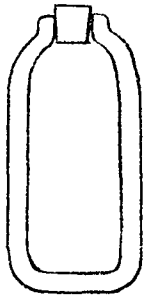
a) Y will rise.
b) X will remain at rest.
c) Y will move toward X.
d) X will fall and move toward Y.

24. Telephone wires are hung as shown. The distance d is a minimum during



a) winter.
b) spring.
c) summer.
d) none of these seasons.

25. The removal of air from between the walls of a thermos bottle would reduce heat loss mainly due to

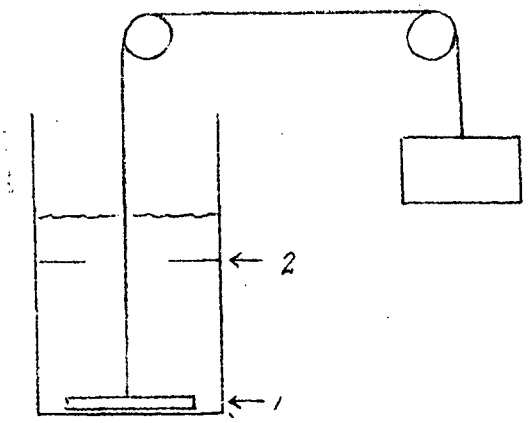


a) conduction and radiation.
b) convection and conduction.
c) radiation and convection.
d) conduction, convection, and radiation.

26. Water at 80° C. is

a) twice as hot as water at 40° C.
b) eighty times as hot as water at 0° C.
c) two units hotter than water at 40° C.
d) eighty units hotter than water at 0° C.

27. A heavy block, attached to a disc, is allowed to fall pulling the disc through the water from position 1 to position 2. Which of the following statements is correct?



a) The temperature of the water rises.
b) The temperature of the water falls.
c) The temperature of the water remains constant.
d) The disc does no work on the water.

28. Ethylene glycol dissolves in water in beaker 1 while table salt dissolves in water in beaker 2.

a) Both mixture 1 and mixture 2 freeze below 0° C.
b) Neither mixture 1 nor mixture 2 freezes below 0° C.
c) Only mixture 1 freezes below 0° C.
d) Only mixture 2 freezes below 0° C.

29. The amount of energy stored in one gram of steam at 100° C. is

a) greater than the amount of energy stored in one gram of water at 100° C.
b) less than the amount of energy stored in one gram of water at 100° C.
c) the same as the amount of energy stored in one gram of ice at 0° C.
d) the same as the amount of energy stored in one gram of water at 100° C.

30. Several drops of water are placed on the bulb of a mercury thermometer and are allowed to evaporate. The mercury level in the thermometer drops because the evaporation of the water

a) leaves the air around the bulb cooler than the surrounding air.
b) removes heat from the region around the bulb.
c) removes heat from the bulb.
d) transfers heat energy to the mercury.

31. Two grams of water at 30 ° C. are poured onto a five gram piece of glass at 0 ° C. The final temperature of water and glass is 20 ° C. The specific heat of glass is

a) .1 $\frac{\text{cal.}}{\text{gm.}}$ per ° C.
b) .2 $\frac{\text{cal.}}{\text{gm.}}$ per ° C.
c) .4 $\frac{\text{cal.}}{\text{gm.}}$ per ° C.
d) .5 $\frac{\text{cal.}}{\text{gm.}}$ per ° C.

32. Four grams of boiling water is poured into twenty grams of water at an unknown temperature. If the final temperature (after mixing) is 60 ° C., what was the unknown temperature?

a) 68 ° C.
b) 52 ° C.
c) 48 ° C.
d) 40 ° C.

APPENDIX D

HYPOTHESIS I

Control Group				Experimental Group			
Loss Score	Orig. Knowl. Score	Read Test Score	N.A. Test Score	Loss Score	Orig. Knowl. Score	Read Test Score	N.A. Test Score
2	19	46	28	0	28	40	17
-4	30	52	29	0	33	41	29
-1	37	70	40	0	14	36	20
-4	31	52	37	3	33	63	35
2	32	56	38	-4	26	45	38
-3	32	55	34	-4	26	45	28
-2	28	48	38	0	18	42	24
-3	17	49	34	-2	26	47	23
-1	21	44	16	-1	24	47	23
7	27	45	34	2	33	49	34
6	31	51	30	4	20	41	31
-2	29	52	30	2	29	50	36
1	26	43	37	7	23	49	31
-1	29	51	35	0	15	18	27
-1	35	61	38	-2	25	37	27
2	23	46	34	-2	23	37	25
1	27	33	33	0	30	48	29
0	37	51	34	0	32	39	28
0	17	39	31	-6	19	47	32
-1	36	68	37	-2	33	41	29
0	31	55	38	3	26	42	38
-4	31	56	35	-3	24	54	31
-7	20	35	21	-4	32	37	22
4	30	53	34	5	27	37	24
1	27	53	30	1	34	60	36
4	26	37	38	-2	25	40	36
0	30	57	35	1	36	57	36
0	33	54	30	-1	19	37	27
0	32	54	32	6	27	41	23
1	33	57	33	1	25	27	29
4	31	39	25	7	21	18	16
-3	30	32	34	1	27	46	23
8	27	39	28	-5	28	35	25
-2	18	45	27				
1	18	40	27				
-1	21	46	31				
-1	16	40	23				
-5	16	36	18				

HYPOTHESIS II

Control Group

Loss Score	Orig. App. Score	Read Test Score	N.A. Test Score
-2	9	46	28
8	22	52	29
0	27	70	40
-2	19	52	37
-5	17	56	38
1	26	55	34
3	17	48	38
-3	15	49	34
7	15	44	16
-4	14	45	34
-1	14	51	30
1	16	52	30
2	18	43	37
0	18	51	35
-3	21	61	38
1	14	46	34
-2	10	33	33
5	24	51	34
3	17	39	31
-2	23	68	37
-5	18	55	38
-2	23	56	35
2	16	35	21
-1	18	53	34
1	23	53	30
-1	19	37	38
-1	18	57	35
-3	22	54	30
5	23	54	32
-3	18	57	33
0	19	32	34
3	11	39	28
6	12	45	27
-3	10	40	27
-4	14	46	31
-3	9	40	23
3	13	36	18

Experimental Group

Loss Score	Orig. App. Score	Read Test Score	N.A. Test Score
-1	16	40	17
1	16	41	29
1	9	36	20
-3	25	63	35
-6	16	45	38
0	18	45	28
-5	10	42	24
1	15	47	23
-3	17	47	23
-5	16	49	34
4	17	50	36
3	16	49	31
-1	11	18	27
2	15	37	27
0	13	37	25
6	23	48	29
2	20	39	28
-2	13	47	32
-4	21	41	29
-3	13	42	38
-5	19	54	31
1	18	37	22
0	15	37	24
1	24	60	36
-7	10	40	36
-5	18	57	36
3	12	37	27
-2	14	41	23
-5	14	27	29
-5	9	18	16
-1	16	46	23
-5	13	35	25

HYPOTHESIS III

DISTRIBUTION OF SUBTEST LOSS SCORES

Differential Loss Scores	Frequency	
	Control Group	Experimental Group
-8		1
-7	1	
-6	1	1
-5	2	1
-4	2	
-3	6	3
-2	3	5
-1	4	4
0	3	7
1		5
2	4	1
3	3	1
4	2	2
5	2	1
6	3	
7	1	
N =	37	32

APPENDIX E

THE DERIVATION OF THE TEST STATISTIC USED IN THE DIFFERENTIAL RETENTION HYPOTHESIS

The Knowledge and Application subtest scores are assumed to be normally distributed with means and variances as shown.

Assume: $X_{Y_z} \sim N\left(\mu_{X_{Y_z}}, \sigma^2_{X_{Y_z}}\right)$

where: X = A or K

Y = E or C

z = 1 or 2

and A represents an Application subtest score

K represents a Knowledge subtest score

E represents the experimental group

C represents the control group

1 represents the initial subtest result

2 represents the final subtest or retest result

Consider $\dfrac{A_{E_2} - A_{E_1}}{\sigma_{A_{E_2} - A_{E_1}}}$, which is the experimental group individual Application subtest difference scaled to its own standard deviation. The denominator in this term is the standard error of Application subtest score difference.

Now, $\left(A_{E_2} - A_{E_1}\right) \sim N\left(\mu_{A_{E_2}} - \mu_{A_{E_1}}, \sigma^2_{A_{E_1}} + \sigma^2_{A_{E_2}} - 2\rho_{A_{E_{1,2}}}\sigma_{A_{E_1}}\sigma_{A_{E_2}}\right)$

where: $\rho_{A_{E_{1,2}}}$ represents the Application subtest-retest population correlation coefficient,

$\sigma^2_{A_{E_1}}$ represents the experimental population initial Application subtest score variance,

$\sigma^2_{A_{E_2}}$ represents the experimental population final Application retest score variance,

$\mu_{A_{E_1}}$ represents the experimental population initial Application subtest mean, and

$\mu_{A_{E_2}}$ represents the experimental population final Application retest mean.

Then, $$\left\{\frac{A_{E_2} - A_{E_1}}{\sigma_{A_{E_2} - A_{E_1}}}\right\} \sim N\left(\frac{\mu_{A_{E_2}} - \mu_{A_{E_1}}}{\sigma_{A_{E_2} - A_{E_1}}}, 1\right).$$

It follows that,

$$\left\{\frac{\bar{A}_{E_2} - \bar{A}_{E_1}}{\sigma_{A_{E_2} - A_{E_1}}}\right\} \sim N\left(\frac{\mu_{A_{E_2}} - \mu_{A_{E_1}}}{\sigma_{A_{E_2} - A_{E_1}}}, \frac{1}{N_E}\right)$$

where: $\bar{A}_{E_1}$ represents the experimental group initial Application subtest mean,

$\bar{A}_{E_2}$ represents the experimental group final Application retest mean, and

N_E represents the size of the experimental group.

Similarly, $$\left\{\frac{\bar{K}_{E_2} - \bar{K}_{E_1}}{\sigma_{K_{E_2} - K_{E_1}}}\right\} \sim N\left(\frac{\mu_{K_{E_2}} - \mu_{K_{E_1}}}{\sigma_{K_{E_2} - K_{E_1}}}, \frac{1}{N_E}\right).$$

Therefore, for the experimental group,

$$\left\{\frac{\bar{A}_{E_2} - \bar{A}_{E_1}}{\sigma_{A_{E_2} - A_{E_1}}} - \frac{\bar{K}_{E_2} - \bar{K}_{E_1}}{\sigma_{K_{E_2} - K_{E_1}}}\right\} \sim$$

$$N\left(\frac{\mu_{A_{E_2}}-\mu_{A_{E_1}}}{\sigma_{A_{E_2}-A_{E_1}}}-\frac{\mu_{K_{E_2}}-\mu_{K_{E_1}}}{\sigma_{K_{E_2}-K_{E_1}}}\ ,\ \frac{1}{N_E}+\frac{1}{N_E}-2\rho_{AK_E}\frac{1}{\sqrt{N_E}}\cdot\frac{1}{\sqrt{N_E}}\right)$$

or $$N\left(\frac{\mu_{A_{E_2}}-\mu_{A_{E_1}}}{\sigma_{A_{E_2}-A_{E_1}}}-\frac{\mu_{K_{E_2}}-\mu_{K_{E_1}}}{\sigma_{K_{E_2}-K_{E_1}}}\ ,\ \frac{2}{N_E}\cdot\left[1-\rho_{AK_E}\right]\right)$$ or

simplifying symbols $N(D_E\ ,\ V_E)$ where ρ_{AK_E} represents the experimental group Knowledge - Application loss score correlation coefficient.

Similarly, for the control group

$$\left\{\frac{\bar{A}_{C_2}-\bar{A}_{C_1}}{\sigma_{A_{C_2}-A_{C_1}}}-\frac{\bar{K}_{C_2}-\bar{K}_{C_1}}{\sigma_{K_{C_2}-K_{C_1}}}\right\}$$ is distributed as

$$N\left(\frac{\mu_{A_{C_2}}-\mu_{A_{C_1}}}{\sigma_{A_{C_2}-A_{C_1}}}-\frac{\mu_{K_{C_2}}-\mu_{K_{C_1}}}{\sigma_{K_{C_2}-K_{C_1}}}\ ,\ \frac{2}{N_C}\cdot\left[1-\rho_{AK_C}\right]\right)$$ or

simplifying symbols $N(D_C\ ,\ V_C)$

Thus the difference between experimental and control group scaled mean differential loss scores,

$$\left\{\left[\frac{\bar{A}_{E_2}-\bar{A}_{E_1}}{\sigma_{A_{E_2}-A_{E_1}}}-\frac{\bar{K}_{E_2}-\bar{K}_{E_1}}{\sigma_{K_{E_2}-K_{E_1}}}\right]-\left[\frac{\bar{A}_{C_2}-\bar{A}_{C_1}}{\sigma_{A_{C_2}-A_{C_1}}}-\frac{\bar{K}_{C_2}-\bar{K}_{C_1}}{\sigma_{K_{C_2}-K_{C_1}}}\right]\right\},$$

is distributed as $N(D_E-D_C\ ,\ V_E+V_C-2\rho\sqrt{V_E}\sqrt{V_C})$.

However, $\rho = 0$ since the experimental and control groups are independent and $D_E = D_C$ under the null hypothesis.

Thus, the test statistic,

$$\frac{\left[\dfrac{\bar{A}_{E_2} - \bar{A}_{E_1}}{\sigma_{A_{E_2} - A_{E_1}}} - \dfrac{\bar{K}_{E_2} - \bar{K}_{E_1}}{\sigma_{K_{E_2} - K_{E_1}}}\right] - \left[\dfrac{\bar{A}_{C_2} - \bar{A}_{C_1}}{\sigma_{A_{C_2} - A_{C_1}}} - \dfrac{\bar{K}_{C_2} - \bar{K}_{C_1}}{\sigma_{K_{C_2} - K_{C_1}}}\right]}{\sqrt{2\left[\dfrac{1 - \rho_{AK_E}}{N_E} + \dfrac{1 - \rho_{AK_C}}{N_C}\right]}}$$

is distributed as $N(0, 1)$ under H_o .