

ABSTRACT

THE TEACHING AND ASSESSMENT OF COGNITIVE STRUCTURE THROUGH THE DIAGRAMMATIC REPRESENTATION OF STRUCTURES OF KNOWLEDGE

By

Jean L. Dyer

One of the basic learning problems in education is how the structure of a discipline can best be transmitted to the cognitive structure of each student. In the present study three aspects of this problem were investigated. First, structure of knowledge was operationally defined. Second, a test which systematically examined the cognitive structures of students was developed. Third, the effects of different modes of presenting structure of knowledge were compared.

Structure of knowledge was defined as the organization of a given area of knowledge, where organization referred to the relationships between elements within that area. Elements were defined as the basic unit within a structure with the type of elements in any structure depending upon the subject matter area itself. Some typical elements are concepts, principles, events, and facts. Relationships were defined as the connections between elements. The following classification of relationships was developed: descriptive, causation, multiple causation, temporal, logical, quantitative,

functional, and composite (an interaction of any of the preceding). Diagrams (Venn diagrams, tables or matrices, lists, and graphs) were used to represent the structure of any area. All structures could be represented by graphs (graph theory), for there is a one-to-one correspondence between graphs and the definition of cognitive structure with points representing elements and lines representing relationships.

Cognitive structure was defined as an individual's organization of knowledge in a certain subject matter area at a given time, where organization again referred to relationships between elements. However, the knowledge structure that the individual has acquired may not coincide with the structure of the subject matter. Using diagrams as representations of structure of knowledge, a procedure for constructing tests which systematically tested each of the specified structural relationships was presented. Transfer structures, covering material that was logically implied by two or more given structures, were also explained. Items were scored to reflect the number of relationships correctly understood. In many cases patterns of item responses were scored rather than using the traditional method of scoring each item independently.

Two different modes of representing structure of knowledge were examined, verbal and diagrammatic. It was expected that diagrams would result in higher performance on acquisition and retention because diagrams would serve as "perceptual blueprints"; separating relevant from irrelevant structural relationships more clearly than verbal statements, organizing material

during acquisition and retention, representing material in a rather stable form for storage, and aiding retrieval of information.

Three treatments were used to present the structure of knowledge in a 7,000 word passage on reliability: a diagram (D), a verbal (V), and a non-review (NR) treatment. The D treatment presented diagrams representing the reliability structure in the following manner: 19 small diagrams, six substructure diagrams which integrated these smaller diagrams, and a diagram which connected all of these substructure diagrams, i.e., the total structure of the passage. These diagrams were placed within the passage following material relevant to the comprehension of each diagram. The V treatment was identical to the D treatment except that the three levels of structural representations were in verbal rather than diagram form. The NR treatment consisted of the reliability passage without either the verbal or the diagram representations of structure. All Ss were given a diagram interpretation program before the administration of the treatment passage. After the passage Ss were given a test over the reliability content. Ss ($n = 156$) were randomly assigned by sex to the three treatments. A control group ($n = 78$), which did not receive any treatment but took the test on the reliability passage, was also used in making some comparisons. The three major dependent variables were performances on a test covering all structural relationships, on a test covering transfer relationships based upon structure, and on a typical multiple-choice achievement test. One week later Ss were again given the tests and

also given a questionnaire pertaining to the experimental materials.

The central hypothesis of the study was that the D treatment would facilitate learning of the structural and transfer relationships on both acquisition and retention more than either the V or NR treatments. However, no differences were expected among the treatment conditions for the typical achievement test. A retention drop for all dependent variables across treatments was expected. Performance on certain sequential dependencies among the structural relationships was expected to be highest for the D treatment. Certain correlations were also expected between some variables; in particular, time spent reading the reliability passage and the major dependent variables, structure and transfer, and the dependent variables on acquisition and retention. The primary exploratory question involved the type of structural relationships which were easy or difficult for the Ss.

The major dependent variables showed no significant differences among the experimental treatments. Except for transfer, the control group's performance was lower than the experimental treatment's performance. Time spent reading the reliability passage was greatest for the D treatment and least for the NR treatment. Since the prerequisite item for all the sequential dependencies was not passed by any S, an analysis of these patterns was not possible. Time was not correlated with the major dependent variables and structure was more highly related to achievement than to transfer. However, achievement, structure, and transfer were correlated for acquisition and retention.

A structural analysis of the Ss structural relationships indicated that certain substructures were more difficult than others, and that all Ss had difficulty with certain types of relationships. In particular, precision in definitions was lacking and causal relationships were confused or incomplete. Knowledge of transfer relationships was consistent with the knowledge of structural relationships. The more difficult substructures often resulted in less consistency in Ss' cognitive structures from acquisition to retention.

The unexpected similarity among experimental treatments was explained by inadequate comprehension of the reliability passage with one reading (as indicated by questionnaire data and absolute performance levels), and by presentation of diagrams too soon in the learning process. Because of inadequate comprehension the underlying theoretical position was not adequately tested. Performance on the items was related to two factors, the chance level of the items (format) and the number of relationships tested by the item (information load). These variables were used to explain correlation patterns and performance on items identified in the sequential dependency patterns.

The differences among time spent reading the reliability passage suggested methodological implications for future research of a similar nature. Despite the general negative results of the experimental treatments, the difficulty Ss had with certain structural relationships supported the usefulness of a structural analysis in testing and in diagnosing learning problems. The

Jean L. Dyer

similarity between Ss' comprehension and confusion of structural relationships, despite different versions of the reliability passage, suggested the need for investigating the ability of individuals to understand different types of relationships.

THE TEACHING AND ASSESSMENT OF COGNITIVE STRUCTURE
THROUGH THE DIAGRAMMATIC REPRESENTATION
OF STRUCTURES OF KNOWLEDGE

by

Jean L^{uise} Dyer

A THESIS

Submitted to

Michigan State University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

College of Education

1969

557263
7-3-69

ACKNOWLEDGMENTS

I wish to thank Dr. Lee Shulman, Dr. John Hunter, Dr. Clessen Martin, and Dr. John Wagner for their cooperation and suggestions in developing the study. I am particularly grateful to my chairman, Dr. Shulman, who gave me the freedom to explore some rather complex areas within educational psychology. In addition, I wish to thank Dr. Hunter, who encouraged me in my initial explorations and whose constructive ideas clarified my own thoughts and provided me with the precision to operationalize the basic concepts in the study.

I would also like to thank my friends and associates who helped me in many ways, especially Jay Powell, who willingly discussed many of the technical aspects of the study with me. Of course, the final words of appreciation go to my husband, Fred, who experienced every success and failure of the study with me over three 'long' years. His suggestions and encouragement were a vital part of the success and completion of every aspect of the project.

TABLE OF CONTENTS

Chapter	Page
I. STRUCTURE OF KNOWLEDGE AND COGNITIVE STRUCTURE	1
Structure of Knowledge	2
Representation of Structure of Knowledge with diagrams	9
Cognitive Structure	17
II. TESTING COGNITIVE STRUCTURE	22
Testing Cognitive Structure Relationships	22
Comparison of This Approach with Other Approaches to Representing and Testing Cognitive Structure	39
III. RATIONALE FOR DIAGRAMMATIC PRESENTATION OF STRUCTURE OF KNOWLEDGE	47
The Role of Diagrams in Learning	48
Individual Differences in Understanding Diagrams	54
Organization of Diagrams within a Passage	55
Overall Sequence of a Passage	58
IV. PILOT STUDY, EXPERIMENTAL MATERIALS, AND HYPOTHESES	60
Pilot Study	60
Experimental Materials	61
Tests	66
Additional Pilot Study Results	69
Hypotheses	70
V. PROCEDURE AND RESULTS	76
Procedure	76
Major Results	80
Minor Results	91

Chapter		Page
VI.	ANALYSIS OF THE COMPREHENSION AND RETENTION OF STRUCTURAL RELATIONSHIPS . . .	105
	Structure	106
	Transfer	138
	Summary and Interpretation	152
VII.	DISCUSSION OF MAJOR RESULTS AND CONCLUSIONS	157
	Time	157
	Achievement, Structure, and Transfer	158
	Sequential Dependencies	161
	Relationships among Main Variables	164
	Implications of the Study.	166

LIST OF TABLES

Table	Page
1. Distribution of Subjects	78
2. Means and Standard Deviations for the Main Dependent Variables for Each Treatment	81
3. Correlation Coefficients among the Main Dependent Variables for all Subjects	90
4. Subjects' Rankings of Substructure Difficulty by Treatment and Group	96
5. Correlation Coefficients among Main Dependent, Substructure, Background, and Questionnaire Variables for all Subjects	97
6. Percentage of Subjects with Perfect Substructures on Acquisition and Retention	107
7. Substructure 1: Acquisition-Retention and Consistency Percentages for Structures 20,21 , and 22a	108
8. Substructure 2: Acquisition-Retention and Consistency Percentages for Structure 2	111
9. Substructure 3: Acquisition-Retention Percentages for Structure 6 - Memory Item	116
10. Substructure 3: Consistency Percentages for Structure 6	122
11. Substructure 3: Acquisition-Retention Percentages for Structures 4 and 5	123
12. Substructure 3: Consistency Percentages for Structures 4 and 5	126
13. Substructure 4: Acquisition-Retention and Consistency Percentages for all Items	128

Table		Page
14.	Substructure 5: Acquisition-Retention Percentages for Structure 19 - Parallel Forms and Tests	133
15.	Substructure 5: Consistency Percentages for Structure 19 - Parallel Forms and Tests	135
16.	Substructure 6: Acquisition-Retention and Consistency Percentages for Structure 7-1st	137
17.	Transfer, Substructure 1: Acquisition-Retention and Consistency Percentages for Structure 22b	139
18.	Transfer, Substructure 2: Acquisition-Retention and Consistency Percentages for Structure 3	140
19.	Transfer, Substructure 5: Acquisition-Retention and Consistency Percentages for Structure 17	141
20.	Transfer, Substructures 3 and 6: Acquisition-Retention Percentages for Structure 7-2nd	143
21.	Transfer, Substructures 3 and 6: Consistency Percentages for Structure 7-2nd	145
22.	Transfer, Substructures 3 and 6: Acquisition-Retention Percentages for Structure 8	147
23.	Transfer, Substructures 3 and 6: Consistency Percentages for Structure 8	149
24.	Transfer, Substructures 3 and 6: Acquisition-Retention Percentages for Structures 9, 10, 11, and 12	150
25.	Transfer, Substructures 3 and 6: Consistency Percentages for Structures 9, 10, 11, and 12	151
26.	Guttman Reproducibility Coefficients on Structure and Transfer	162
27.	Rank Correlations: Actual Difficulty with Format and Information on Structure and Transfer	163

Table		Page
28.	Analysis of Variance for Time and Errors	295
29.	Scheffé' Multiple Comparisons on Time	296
30.	Analysis of Variance for Achievement	297
31.	Analysis of Covariance for Achievement	298
32.	Scheffé' Multiple Comparisons on Achievement for Six Treatments and Control	299
33.	Analysis of Variance for Structure	300
34.	Analysis of Covariance for Structure	301
35.	Scheffé' Multiple Comparisons on Structure for Six Treatments and Control	302
36.	Analysis of Variance for Transfer	303
37.	Analysis of Covariance for Transfer	304
38.	Analysis of Variance for Substructure on Acquisition - Six Treatments and Control	305
39.	Analysis of Variance for Substructures on Retention - Six Treatments	306
40.	Scheffé' Multiple Comparisons on Substructure- Acquisition for Six Treatments and Control	307
41.	Means and Standard Deviations on Substructure- Retention for Six Treatments	309
42.	Responses to the Diagram and Verbal Questionnaire Items	310
43.	Correlations among Aptitude Scores and Main Dependent Variables	312

LIST OF APPENDICES

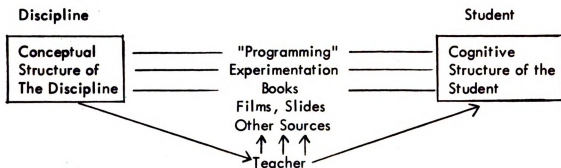
Appendix	Page
A. Pilot Questionnaire	176
B. Treatment Questionnaire	180
C. Diagram Interpretation Program	183
D. Outline of Reliability Passage	215
E. Reliability Passage	216
F. Diagrams and Corresponding Verbal Statements	245
G. Test and Test Analysis	260
H. Order of Test Items	292
I. Guttman Dependencies	293
J. Analysis of Variance for Time and Errors	295
K. Analysis of Variance and Covariance for Achievement	297
L. Analysis of Variance and Covariance for Structure	300
M. Analysis of Variance and Covariance for Transfer	303
N. Analysis of Variance for Substructures	305
O. Questionnaire Items Unique to Diagram and Verbal Treatments	310
P. Aptitude Correlations	312

CHAPTER I

STRUCTURE OF KNOWLEDGE AND COGNITIVE STRUCTURE

Many educational psychologists use rather broad and vague concepts which lack the precision necessary for fruitful application to classroom learning. Two rather common, yet vague, concepts in the literature today are structure of knowledge and cognitive structure. Cognitive structure refers to what a person knows, whereas structure of knowledge refers to what experts have decided he should know. Diagramming is proposed as a method of clarifying these two concepts, since it provides a method of representing structures of knowledge and a model for testing an individual's cognitive structure. In addition, diagrams provide new approaches to curriculum development and to the assessment of learning.

Novak (1966, p. 249-253) presented the following model of the educational process whereby the conceptual structure of a given discipline is transmitted to the student by various media such as books, teachers, and films with the student storing and integrating this information within his own cognitive structure.



The purpose of this study was to investigate all three aspects of this model. It attempted (1) to define conceptual structure of a discipline (structure of knowledge) and to develop a method of representing this structure, (2) to develop a test specifically of cognitive structure, and (3) to determine the effects upon cognitive structure of various modes of presenting structure of knowledge.

Structure of Knowledge

Many of the new social studies and mathematics curricula emphasize structure of knowledge. This trend is partly the result of Bruner's (1963) stress upon designing curricula that reflected the basic structure of a field of knowledge. According to Bruner (1963): "to learn structure is to learn how things are related (p.7)" . . . "it implies learning the underlying principles, attitudes, and/or regularities of a subject (Bruner, 1966a, p.249)." Structure was conceived as the most economical representation of a discipline; namely, the rules or propositions which generate it (Bruner, 1966b, p. 201-203).

Morrisett (1967) defined structure of knowledge similarly;

the arrangement and interrelationships of parts within a whole. A structure can refer to the relationship of concepts to each other; for example the concepts, "economic system" and "political system" may be related to each other in a structure called a "social system." Conversely, a concept may itself have a structure. The concept "economic system" can also be thought of as a structure having component concepts such as "money" and "spending" which are structurally related to each other (p. 4)

By citing "economic system" and "social system" as structures, where economic system was part of the structure called a social system, Morrisett showed that structures could have different levels of abstraction and/or complexity. But he did not clarify the meanings of interrelationships (or arrangement) and parts. For example, he distinguished theory from structure, a theory being

a general statement about relationships among facts, where these facts have been organized into concepts. A theory is a structure of concepts; it states a relationship - often a casual relationship among the concepts (p. 5).

In essence Morrisett defined a theory as a specific type of structure, with a specific type of part (facts organized into concepts) and a specific relationship (casual). However, later he treated structure and theory as two separate entities within any curriculum, theory implying more than just a structure. It seemed that the meaning of structure of knowledge needed to be clarified by defining what is meant by "part" and what is meant by "arrangement" or "interrelationships".

The present definition of structure of knowledge represents an

attempt to clarify and expand upon Bruner's and Morrisett's approaches.

Structure of knowledge is defined as the organization of a given area of knowledge, where organization refers to the relationships between elements within that area.

Elements

In this definition, element might mean what is usually termed a concept, a principle, an event, a fact, an object, a theory, or a sub-structure, which could be larger or smaller than a theory. The type of element in a structure depends upon the particular subject matter area. The criterion for determining the elements of a given structure is subjective, not yet based upon a mathematical or experimental procedure.

The lowest level of an element is an event, fact, or object, where the elements are not abstractions, e.g., President Johnson talked with Premier Kosygin on June 25, 1967. Concepts, being abstract, are at a higher level. Some examples of basic concepts from statistics would be mean, variance, and correlation; from science, energy, matter, neutron, and electron. Principles are at the next level of complexity. Some principles in physics would be the laws of magnetism, the gas laws, and buoyancy principles.

Gagné's (1966) distinction between concepts and principles is used. For Gagné concepts refer essentially to equivalence classes of objects or object-qualities. One type of concept, concept by observation, is learned

through observation of positive and negative instances and another type, concept by definition, is learned through verbal communication. Principles, on the other hand, are "composed of two or more concepts having an ordered relationship between them (Gagné, 1966, p. 98)".

But this distinction is not always clear as Gagné himself pointed out in referring to concepts by definition,

The other [type] is a concept by definition, which is in a formal sense the same as a principle. It is a combination of simpler concepts and is typically learned by human beings via verbal statements that provide the cues to recall of component concepts and to their correct ordering (p. 90).

This distinction between concepts by definition and principles is particularly unclear when the mode of presentation is only that of written material, which is the way many concepts are learned, as Gagné pointed out.

Gagné (1966) made two other major distinctions between concepts and principles. Gagné stated that the behavioral criteria for knowing concepts and principles were different; for a concept it is identification while for a principle it is demonstration. Concept testing involves a choice from a number of alternatives; principle testing involves a situation where performance reflects identification of component concepts and the operation relating them to each other. Considering the previously stated similarity between concepts by definition and principles, this seems to be an arbitrary criterion. The other major distinction was that of mediation; a concept representing a single mediator and a principle representing a sequence of

mediators. However, this criterion is also unsatisfactory for complex concepts could easily involve a sequence of mediators. Even though Gagné's distinction between concepts and principles was inadequate his classification of concepts into two types, concepts by observation and concepts by definition was used.

Returning to the original discussion of structure and the term "element", a certain structure might be an explanation of a concept or principle. Other structures might involve comparison of principles or concepts, implications of principles or concepts, combination or transformation of principles into more inclusive ones, etc.

When elements are substructures such as theories, the structure or organization involves a hierarchical sequence of interrelationships. There could also be many generalizations which form substructures not complex enough to be called theories and even other substructures more complex than theories.

Relationships

Relationships are the connections between elements in a body of knowledge. The relationships given below are intended to be exhaustive of the entire class of relationships. However this may not be the case.

Descriptive:	is most apparent in definitions or characterizations of things, e.g., characteristics of types of rocks, criteria of good pop art,
--------------	--

advantages and disadvantages of two theoretical positions.

Causation:	of the form "A causes B"
Multiple Causation:	of the form "(A,B,C, . .) causes (W,X,Y, . .)"
Temporal:	of the form "A precedes B"
Logical:	refers to the logical connectives - and, or, not, if then, etc. One important subset here is the subset-set or inclusion-exclusion relationship, e.g., classification of rocks, trees, or mammals; types of statistical tests, etc.
Quantitative:	mathematical relationships such as equality, inequality, proportionality, mathematical functions, order, addition, etc.
Functional:	excludes mathematical functions. Refers to relationships that express purpose, use, action, direction, transformation, etc., e.g., a computer processes data, a skillet is for cooking food, Edmonton, Alberta is northwest of East Lansing, Michigan, John is the father of Bill, etc.
Composite:	any complex interaction of the relationships listed above, e.g., temporal-causal.

A closer examination of different subject matter areas shows that within and across areas different relationships may exist. Some concepts have a subset-set type of relationship, e.g., in matrix theory the identity matrix is a subset of the set of all diagonal matrices and in history the set of explorations to America could be divided into subsets according to the country for which the explorer sailed, to the period of time, or perhaps by geographical

areas of exploration on the mainland. In some areas terms are related in the sense that one describes the other; a psychological learning theory might be described as behaviorist, neo-behaviorist, or cognitive and in geometry triangles can be described as equiangular, equilateral, similar, congruent, etc. Another common relationship is that of logical implication, e.g., an equilateral triangle logically implies an equiangular triangle and vice versa, and much unsystematic or error variance in test scores implies low reliability.

Range of Application of the Present Definition

This definition of structure of knowledge is quite broad, including structures which are simple, only two elements and one relationship, to complex structures involving many elements and many relationships including entire disciplines. Structures can also vary in degree of abstraction of elements, i.e., from events or facts to theories Bourne (1966) reviewed concept learning studies which varied the type of rule (relationship) that combined the defining attributes of a concept, e.g., conjunctive, disjunctive, relational, joint denial, conditional, etc. Rules showed different degrees of difficulty. These results would imply that cognitive structure relationships also vary in degree of difficulty, reflecting perhaps differences in degree of abstraction and/or complexity.

The present definition is not limited to what might be called the "basic" structure of a discipline where basic refers to the most important

ideas, the underlying principles, and/or regularities of a discipline. Since the basic structure of any discipline is to a certain extent determined by experts in that field itself, different structures could be identified, distinguished primarily by a choice of different elements and in some cases by different relationships among these elements. But given a passage of material encompassing a smaller content area than an entire discipline, it is assumed that the same structure would be identified by different individuals.

Representation of Structure of Knowledge with Diagrams

The preceding definition and explanation of structure of knowledge implies that these structures can be represented by diagrams which illustrate the elements and their relationships. In other words, the structure of a subject matter or of a passage can be "seen", just as the structure of a crystal or of a building can be "seen."

Diagrams or graphics may be classified into four types:

Venn diagrams, tables or matrices ($n \times m$ array), lists, and graphs, that is the graphs of graph theory (Harary, Norman, and Cartwright, 1965). This classification can represent all of the previously listed relationships. In fact, even by using only graphs all structures could be represented, simply by letting points represent elements and lines represent relationships. Berlyne (1965) in discussing situational thoughts and transformational thoughts, which

are similar to cognitive elements and relationships respectively, has also stated that graph theory could represent all structures. For Berlyne "a node [a point] can stand for a situational thought and a branch [connecting line] for a transformational thought leading from one situational thought to another (p. 200)." A more extensive analysis of Berlyne's approach is given later.

Generally types of diagrams other than graphs are used because some diagrams lend themselves well to representation of particular relationships. A useful correspondence between diagrams and relationships is the following:

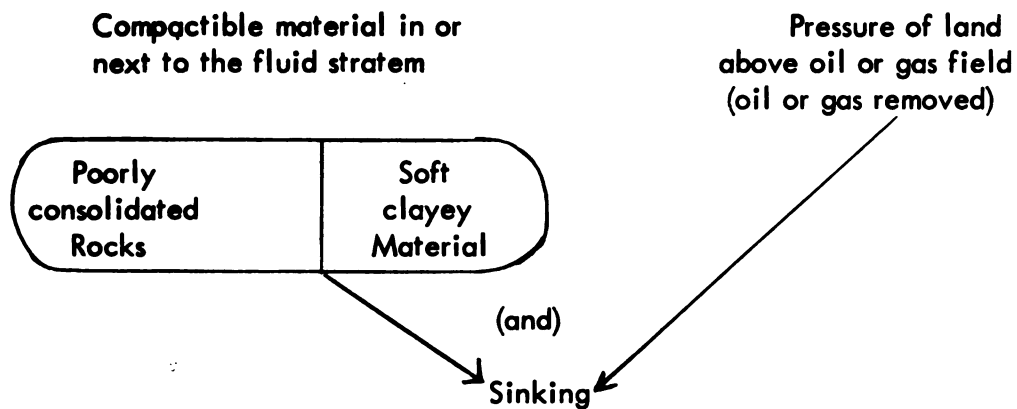
Venn diagrams:	subset-set
Tables or Matrices:	descriptive, causal, logical, multiple causation
Lists:	subset-set, descriptive (outline is one type of list)
Graphs:	causal, multiple causation, logical, quantitative, functional, composite (time lines in history and flow charts in chemistry are specific examples of graphs)

A series of examples will illustrate how a diagram could represent the structure of a given area. Below is a paragraph on sinking of land followed by a diagram representing the structure of that passage.

Geological subsidence or sinking of lands results from tapping the earth for oil or gas. Near Long Beach, California the land above the Wilmington oil field sank until it had become a bowl up

to 26 feet deep over an area of 22 square miles. The slow subsidence of land ruined buildings, racked pavements, twisted railroad tracks, and wrecked bridges.

The explanation for such phenomenon is as follows. Liquid or gas is generally drawn from a stratum of porous rock whose pores are filled with the fluid under pressure. If the rock is well consolidated (if its grains are well cemented together) it will usually continue to support the weight of the rock and earth on top after the fluid is withdrawn. However, if the fluid-holding rock is a poorly consolidated, easily-molded sandstone, once the supporting pressure of the fluid has been withdrawn from its pores, the pressure of the overburden compacts the rock, and the ground above subsides by the amount by which the rock is compressed. Other factors besides the mechanical strength of the fluid-containing rock may contribute to subsidence. For example, subsidence is more likely if soft, clayey material (which is easily compacted) is present in or next to the fluid stratum.



This structure is represented by a combination of a Venn diagram and a graph, representing subset-set and multiple causation relationships respectively. The subset-set relationship is the two types of compactable material (poorly consolidated rocks and soft, clayey material) in or next to the fluid stratum. The subsidence or sinking of land is the joint effect of compactable material and the pressure of land above the oil or gas field.

In this case the arrow (→) represents "causes".

An example of a matrix or chart is given below.

Types of Rocks			
Igneous	Sedimentary	Metamorphic	
X			Molten rock which has cooled and hardened
	X		Rock grains locked together by pressure and cementing material
		X	Rock with changed mineral content

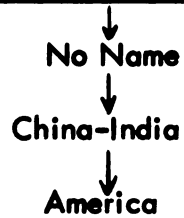
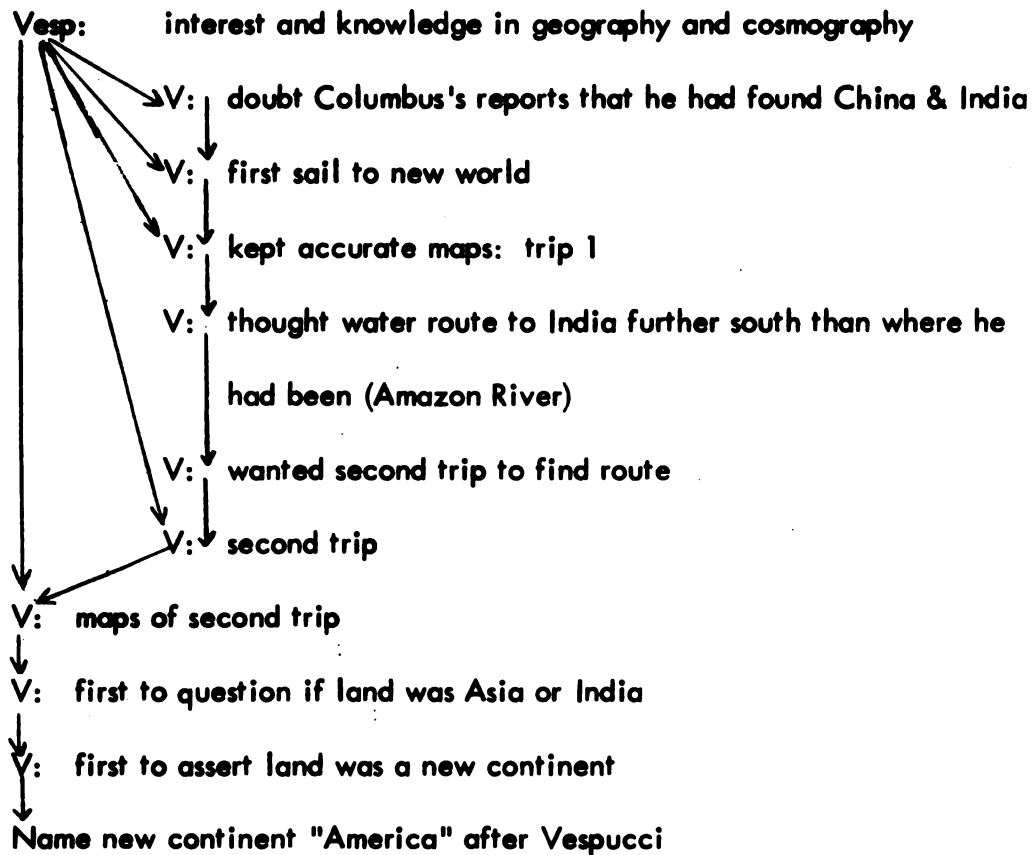
Here the diagram represents the formation description of types of rocks.

Note that the pattern of checks (X) indicates non-overlapping formation processes.

Several structures could be diagrammed and then interrelated.

For example, the following four structures were taken from an article for elementary students justifying why America was named after Amerigo Vespucci rather than Columbus.

Time Line		Location of water route to India	
1492	Vespucci to Spain Columbus 1		Straight west of Europe
1493	Columbus 2	Vespucci:	↓ Further south than West Indies
1499	Vespucci 1	Vespucci:	↓ south of Amazon River
1501	Vespucci 2	Vespucci:	↓ Further south than southern Argentina
1507	Name America		

Name of New WorldCausal Sequence: Why New World named after Vespucci

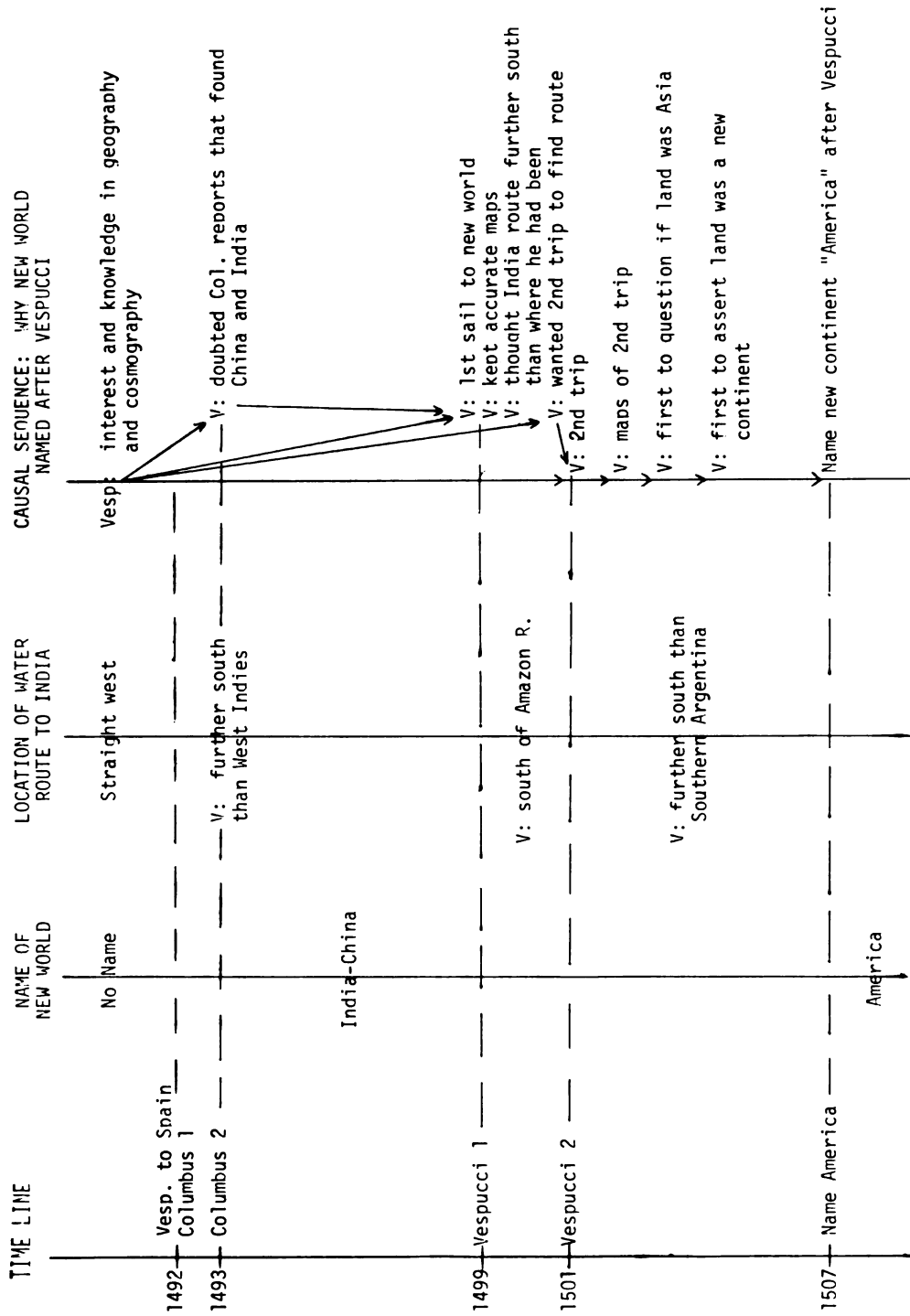
These four substructures are not independent of one another and can be related quite meaningfully on a time dimension. A larger structure is thereby created where the substructures constitute the elements and these are

connected by a temporal relationship (see page 15).

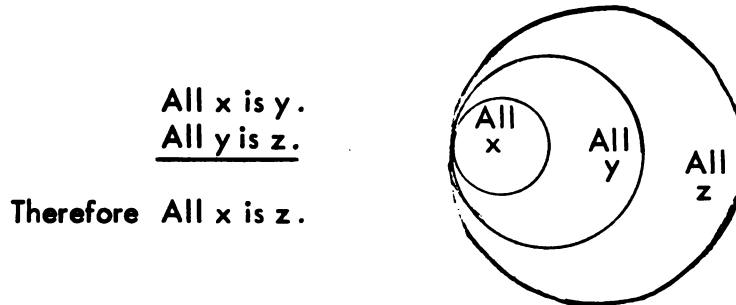
In summary, diagramming a given area requires systematic identification of the basic elements and the relationships among them, followed by selection of an appropriate diagram form to represent the structure. This form of representation can be classified as a symbolic, not an ikonic mode (Bruner, 1964, 1966b, p. 202, 252; Bruner, Oliver, and Greenfield, 1966). For Bruner the ikonic mode refers to summary images that "stand for" the thing represented, where the image must possess parts similar to parts of the object thereby being isomorphic with distinctive features of the thing imaged. The symbolic mode refers to a set of symbolic or logical propositions, where no degree of correspondence with the thing signified is required. Diagrams, because they are not isomorphic with the thing represented, are part of the symbolic mode.

Previous Use of Diagrams to Represent Structure of Knowledge

This methodological procedure of graphically representing structure is not entirely new in the field of educational psychology, although it does not appear to have been applied in a systematic fashion to educational problems. The neo-behaviorist literature has many diagrams illustrating verbal habit family hierarchies, conditioning processes, and mediation. Goss (1961), a neo-behaviorist, referred to $n \times n$ tables and tree diagrams as ways of representing conceptual schemes.

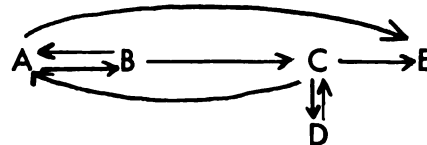


In the cognitive tradition, Hartmann (1942, p. 195) illustrated how a standard syllogistic form could be represented by a Venn diagram.



Bruner (1963, p. 311-312) illustrated how the following list of plane connections between cities could be represented by a simple graph with considerable increase in the economy of presentation.

Boston to Concord
 Danbury to Concord
 Albany to Boston
 Concord to Elmira
 Concord to Danbury
 Albany to Elmira
 Boston to Albany
 Concord to Albany



On a larger scale Senesh (1967) discussed the structure of disciplines of economics, political science, sociology, anthropology, and geography. In all cases he represented or illustrated the respective structure with a diagram, which could be classified as a form of graph. The basic concepts were represented by boxes and the relationships among these concepts represented by lines. In general, the lines did not represent a specific relationship among basic concepts but functional only to illustrate that a

connection or connections existed. Thus Senesh's approach is similar to but not as systematic as the diagram procedure presented here.

Cognitive Structure

Novak's (1966) model included the cognitive structure of the student as well as the structure of the discipline. Cognitive structure is a currently popular, although not new, term used by cognitive theorists in attempting to explain acquisition of knowledge. Ausubel (1963, p. 26) defined cognitive structure as "an individual's organization, stability and clarity of knowledge in a particular subject matter field at any given time." "Organization," "stability" and "clarity" as Ausubel used them appeared to be dimensions by which one could characterize various structures rather than being defining attributes. Ausubel always seemed to be referring to the degree of organization, the degree of stability and the degree of clarity, rather than saying that cognitive structures were organized, clear, and stable. Ausubel's definition is difficult to operationalize because he did not elaborate on the meanings of these terms.

Reitman (1965) attempted to clarify cognitive structure and thought processes through a computer simulation model using the list processing language, IPL-V. His approach to cognitive structure was that "we may regard the whole problem of cognitive structure as a matter of sets of cognitive elements interconnected together in complex networks. Relations

supply the connective tissue that ties the individual sets together into a network (p. 91)." Reitman tended to limit examples of cognitive elements to properties or names of objects, although this limitation was not implied by the term "cognitive element." He stated that a relation could be any connection or linkage among elements ranging from "functional relations (one thing depending upon another) to similarity and equivalence relations (one thing like or unlike another in some respect) and even to systems of interconnections of the sort we find in family relations (p. 96)."

Goss (1961) presented the notion of conceptual schemes which function as mediators organizing particular items or objects. A conceptual scheme was defined as "one or more sets of categories or two or more variables that stand in ordinal, classificatory, or functional [mathematical] relationships to each other (p. 42)." All of these approaches are similar in that they refer to some form of organization or relationship between things variously referred to as cognitive elements, categories, or variables.

Cognitive structure is defined similarly to structure of knowledge. More specifically, it is an individual's organization of knowledge in a certain subject matter area at a given time, where organization refers to the relationships between cognitive elements. Cognitive elements and relationships have been explained previously when discussing the terms "element" and "relationship" as they applied to structure of knowledge. However, cognitive elements and cognitive relationships refer to what the individual

has acquired, which may not agree with the structure of the subject matter. Colloquially speaking, cognitive structure refers to what the person knows, whereas structure of knowledge refers to what experts have decided he should know. An individual's cognitive structure is composed of many inter-related substructures, varying with the content which has been acquired and retained by that individual.

This definition of cognitive structure is more inclusive and flexible than Reitman's and Goss's and reduces the ambiguity of Ausubel's definition. With this definition, "stability" and "clarity" refer to properties of cognitive structure. Goss's (1961) usage of the terms "category" and "variable" is similar in scope to that of "cognitive element," although he referred to functional relationships only in the mathematical sense, whereas the present definition includes additional relationships. Reitman's (1965) concept of relations which tie elements into a network implies the concept of organization as used here. Although he did not specify the precise meaning of elements and relationships, the terms are similar to those in the present definition. As mentioned later Reitman's method of representing cognitive structure differs from the present.

Berlyne's (1965, p. 114-115) distinction between situational thoughts and transformational thoughts is also similar to the distinction between cognitive elements and relationships, although Berlyne was primarily concerned with the role of these two types of thoughts (each

composed of an implicit response and its feedback stimulus) in directed thinking rather than with the product of learning or thinking. A situational thought represents "an external stimulus situation (p. 114)" and a transformational thought represents an operation which changes one situational thought into another. For Berlyne these transformational responses or thoughts are transformations in the mathematical sense as well. For each transformation there is a corresponding domain and range, such that a transformation applied to an element in the domain yields an element in the range. In other words, applying a transformation to a situational thought yields another (or the same, depending upon the transformation) situational thought.

The present approach is similar to Berlyne's. Stating that an individual "knows that A and B are related by X", is another way of saying that he can apply transformation X to A to obtain B. For example, if an individual "knows" that "systematic factors cause systematic variation" then when he applies the transformational thought of "cause of" to the situational thought of "systematic variation" he obtains "systematic factors" and not vice versa. It is also assumed that the individual can operate in a reverse fashion, i.e., he can get from "systematic factors" to "systematic variation" by applying the transformational thought "result of." Additional comparisons with Berlyne's approach are given later.

In this chapter definitions of structure of knowledge and of cognitive structure were presented with an explanation of how diagrams

could be used to represent both types of structures . A comparison with other definitions of structure of knowledge and cognitive structure was also given . In the next chapter a procedure for systematically testing for cognitive structure relationships will be presented and will then be compared with traditional testing procedures .

CHAPTER II

TESTING COGNITIVE STRUCTURE

Having defined structure of knowledge and cognitive structure, a procedure for testing cognitive structure based on these definitions will be presented. It will then be compared with other methodological approaches.

Testing Cognitive Structure Relationships

Defining both structure of knowledge and cognitive structure in terms of relationships between elements and then diagramming the structure of subject matter provides a guide for testing whether a student has learned these relationships and if not, the relationships he has learned. It is not assumed that an individual automatically stores a graph or matrix in his head, although through visual imagery he might use such a form in recall and thought processes. The diagrams which represent the structure of an area serve as blueprints for constructing test items which examine the relationships between elements within an individual's cognitive structure. Knowledge of elements is not directly tested with the present procedure.

In general, multiple choice tests (Ausubel and Fitzgerald, 1961; Ausubel, Robbins and Blake, 1957; Ausubel and Youssef, 1963, 1965; Fitzgerald and Ausubel, 1963) have been used to measure cognitive structure.

The results were then interpreted by using the total score from the test. But the total score as such did not reflect structure, nor the organization, stability and clarity of it. At least no attempts were made to illustrate such relationships. The method of testing for cognitive structure developed in the present study was based on the diagrams which represented the structure of the subject matter. Test items were constructed which systematically examined all of the relationships between elements which were represented in the diagram. Data were scores which reflected the nature of structural relationships attained by the individual. The scores did not reflect simply the number of correct items.

An individual's cognitive structure may not correspond with the structure of a given subject matter area, i.e., he may not have learned and retained the same relationships between elements that exist within the subject matter itself. His cognitive structure may differ from the corresponding knowledge structure in the following ways:

- a. all the structure could be absent
- b. part of the structure (elements and/or relationship) could be present and the rest absent
- c. a different structure could exist, i.e., different relationships and/or different elements present
- d. a combination of b and c could exist

Essay questions could be used to test for cognitive structure, but such items have the undesirable property that students are reluctant to put down ideas because they are afraid the idea is wrong. Yet often these ideas

are the ones in which a researcher is most interested. However, an objective test requires the student to answer each question. Therefore the recognition method was used where the elements were given in the item stem and/or alternative and a correct response indicated knowledge of the proper relationship between elements. Since only relationships are systematically examined this approach has some limitations, reflecting only part of the individual's cognitive structure; but it does insure that all of the structure of knowledge is examined. Such a test can be used to determine how well the structure of the material or passage has been transmitted to the student's cognitive structure. If items are constructed to yield enough information, it is possible to "diagram" an individual's cognitive structure from the test results.

In such a test each item or group of items is analyzed for the number of relationships tested and is then scored accordingly. In the typical achievement test simply because two individuals have the same score on a test does not necessarily mean that they have the same cognitive structure or that they have answered items in the same way. For example, two people could have answered 50 out of 100 items correctly yet have comprehended non-overlapping segments of the material. To a teacher or researcher, inferences for those two individuals would be quite different. Items on a cognitive structure test are often scored on the basis of patterns of responses, different patterns reflecting different cognitive structures rather than scoring each item independently. No attempt has been made to quantify the degree of structure.

A series of examples will show how items testing cognitive structure relationships can be constructed and how the items are analyzed for the number of relationships tested. The concept of reliability has been used in each of these examples. A brief explanation of some of the terms used in the items will be presented to clarify confusions that might be created by the item analysis itself. Reliability refers to the consistency with which a test measures whatever it purports to measure, or the degree to which a test may be depended upon to yield similar test results under similar circumstances. Since no perfectly reliable test exists, variability among scores earned by an individual over repeated testing will occur. The two major types of variation in test scores that can occur are systematic and unsystematic variation (which are caused by systematic and unsystematic factors, respectively). Systematic variation is characterized by a systematic change in scores while unsystematic variation is characterized by random fluctuations in scores. The greater the unsystematic variation then the lower the degree of reliability of a test. Thus in order to increase the reliability of a test, unsystematic factors need to be controlled. In addition, unsystematic factors can be classified into two types, varying and constant. The basic distinction between these two types is that they have different effects upon score variation when tests are given on different occasions.

The quantitative measure of the degree of reliability of a test is called the reliability coefficient. It expresses the extent to which scores on

any type of test can predict scores on a similar type of test . When unsystematic variation is great over a series of similar tests the reliability coefficient is low . Thus the reliability coefficient can be classified as one type of correlation coefficient . There are three ways of estimating the reliability coefficient for a test: (a) test-retest; estimation from the correlation coefficient between scores on repetitions of the same test , (b) parallel forms; estimation from the correlation coefficient between scores on parallel (comparable) forms of a test , and (c) internal consistency; estimation from the correlation coefficient among comparable parts of a test . The series of examples of items given below illustrate how an individual's knowledge of relationships among these basic concepts can be examined .

A simple application type question could test for the relationships among the methods of estimating reliability coefficients and their characteristics (defining attributes) as represented by the following chart . The answers directly reflect if a student has the correct relationships among the methods of estimating reliability coefficients and the defining characteristics of each . Clarity of subjects' cognitive structure is checked by requiring the same response to the first and last parts of the item .

		Test Retest	Internal Consistency	Parallel Forms Immed.	Parallel Forms Delayed
Type of Test Admin.	Identical	X	X		
	Similar			X	X
Time of Testing	Same Occasion		X	X	
	Different Occasion	X			X
Number of Times Test Admin.	Once		X		
	More Than Once	X		X	X

Application Question (i.e., application, because the examples are different from the text)

The following methods are sometimes used for estimating reliability coefficients:

- (a) test-retest
- (b) internal consistency
- (c) parallel forms - immediate
- (d) parallel forms - delayed

For each of the following situations determine which reliability method is described and place that letter (a,b,c,d) on the line preceding the corresponding statement.

_____ The group form of the Stanford-Binet intelligence test was given to the sixth grade class on the opening day of school.

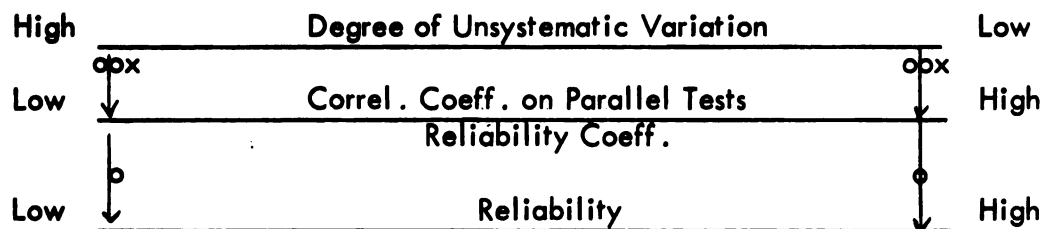
_____ The Iowa Tests of Basic Skills were administered to all transfer students; form A on the first day of school and form B two weeks later.

_____ The teacher gave the same pre and post test on one chapter in the text.

_____ Both forms of a personality scale were administered to a group of nurses upon their graduation.

_____ A final examination was given to all students during the last class period of the day.

Below is an example of scoring a pattern of responses with the corresponding analysis of the patterns, which is in turn related to diagrams representing the structure of a certain area. The small circles (o) in the diagrams indicate the relationships which are tested in the item itself; the correct pattern "tfffft" (a true and f true) reflecting each of these relationships as indicated by the analysis. The small checks (x) are the relationships reflected by the pattern "tfftff" (a true and d true) also indicated by the analysis.



		Reliability	Reliability Coeff.	Correlation Coeff.
Quant. Index		xo	X xo	X xo
Degree of Unsyst. Var. in Scores		X	X	X
Tests	Par.	X	X	X
	Not Par.			X

Question over these two structures:

If we found a low degree of relationship between tests B and C (are parallel tests) we would expect this to be reflected in quantitative indices (index) such as

- a. low correlation coefficient
- b. high correlation coefficient
- c. low reliability
- d. high reliability coefficient
- e. high reliability
- f. low reliability coefficient

Analysis

If a true (b false for consistency) - 3 points

- (1) quantitative index
- (2) two continuums; unsystematic variation and correlation coefficient

If f true (d false for consistency) - 3 points

- (1) quantitative index
- (2) two continuums; unsystematic variation and reliability coefficient

If both a and f true - 1 point - equality of correlation coefficient on parallel tests and reliability coefficient.

Consider c and e combinations when a and f true.

If both c and e false - 3 points

- (1) reliability not a quantitative index
- (2) infer continuums correct - reliability

If c true and e false - 2 points

- (2) continuum correct, but quantitative index wrong

If c and e both true - inconsistent

If c false and e true - no points

If a true and f false. f could be false because of two reasons - continuum or quantitative index. Check by alternative d.

If d marked true is because of continuum, but incorrect, yet has quantitative index - 1 point

If d marked false - no points - can not infer about continuum and quantitative index is wrong.

Consider c and e combinations.

If c true and e false - 2 points for continuum

If c false and e false - 1 point for index

If c false and e true - no points

If a false and f true, a could be false because of either continuum or index.

Check by alternative b.

If b true - 1 point for index (continuum wrong)

If b false - no points (cannot infer about continuum, index wrong)

Consider c and e combinations

If c true and e false - 2 points for continuum

If c false and e false - 1 point for index

If c false and e true - no points

If both a and f marked false, reason is either continuum or quantitative index.

Check by alternative d and b.

If d true - 1 point for index

If d false - no points

If b true - 1 point for index

If b false - no points

C and e combinations follow same pattern as outlined in the previous section.

Using this analysis we then have the following consistent patterns. The remaining patterns are inconsistent, scored -1. (A blank indicates "false".)

[illegible]

If a student responded with the appropriate pattern of responses the diagram of his cognitive structure would be identical to the preceding diagram (relationships indicated with circles only). He would also be given 10 points for such a pattern, it reflecting knowledge of 10 appropriate relationships among elements. If he had responded with "tffftff" we would then examine the situation where a is true and f is false. Under that analysis f could be false for two reasons, which could be checked by examining the response to alternative d. If d is marked true (as is the case) both parts of the alternative must be true from the student's point of view (both continuum and quantitative index). But, in fact, only the quantitative index relationships is correct, therefore allowing for only one point. Then c and e combinations need to be examined. This analysis shows that if such a pattern had been given only five relationships were correct in the student's cognitive structure yielding a score of five points for that pattern. Note that such an analysis gives the number of correct relationships and also tells which of these relationships are appropriate.

There is a broader classification of response patterns that can be used. In general, response patterns can be classified as consistent or

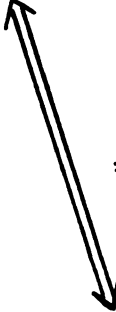
inconsistent. If a pattern is inconsistent a student has contradicted himself. Consistent patterns can be of three types: (a) perfect, meaning that there is an isomorphic relationship between the student's cognitive structure and the structure of the subject matter, (b) incorrect but without contradictions, i.e., the student has answered consistently but wrong throughout the given set of items, and (c) partly correct, i.e., a subset of items has been answered correctly, and the student has not contradicted himself on the other items (answer may reflect an omission of a relationship or confusion of relationships but not a contradiction). An example of a contradiction would be where a student stated that internal consistency and parallel forms-delayed tests yielded the highest reliability estimates and then later stated that internal consistency and test-retest estimates yielded lowest reliability indices. A student would be "confused" if he stated that systematic factors caused unsystematic variation and unsystematic factors caused systematic variation. From both a teacher's and researcher's viewpoint this type of information could be quite valuable, indicating different types of problems for different students.

Testing for Transfer

It is also possible to generate a transfer structure, a structure that is not presented in the material itself, yet is logically implied from what is given. This is done by critically examining the relationship between two or more substructures. Below are two such substructures which yield a transfer structure because of their specific

relationship. Thus transfer questions can be generated which go beyond the material presented and are not merely new applications of an idea or principle.

		System. Factors		Unsystem. Factors		
				Constant	Varying	
Effect	Same	x	x	x		
	Diff.				x	x
Indiv.	Same	x			x	
	Diff.					x
	All		x	x		
Occas.	Same		x	x		x
	Diff.	x			x	



Methods of Estimating Reliability Coefficients					
		Test Retest	Int. Consist.	Paral. Forms Immed.	Paral. Forms Del.
Type of Test	Ident.	x	x		
	Similar			x	x
Time of Testing	Same		x	x	
	Occas.				
	Diff.	x			x
Number Times Test Admin.	Once		x		
	More than Once	x		x	x

Transfer Structure

		Test Retest	Int. Consist.	Paral. Forms Immed.	Paral. Forms Del.
Systematic Factors		x			x
Unsystematic Factors	Constant	x			x
	Varying	x	x	x	x

A transfer question which could be drawn from this structure could be tested.

Transfer Item

Suppose we want to determine the reliability of a newly constructed test called the "Student Happiness Test." Two comparable forms, A and B, were prepared. The decision made to use all four basic methods of estimating reliability coefficients rather than just one or two.

Mark each of the following statements "true" or "false" on the basis of this information. Assume that similar students and group testing were used for all estimation methods.

- a. IC and PFI would give the highest reliability indices.
- b. TR and PFD would give the lowest reliability indices.
- c. IC and PFI would give the lowest reliability indices.
- d. TR and PFC would give the highest reliability indices.
- e. PFD and PFI would give the lowest reliability indices.
- f. IC and TR would give the highest reliability indices.
- g. PFD and PFI would give the highest reliability indices.
- h. IC and TR would give the lowest reliability indices.

This item was constructed on the basis that constant unsystematic factors can not be distinguished when testing is done on the same occasion, thus giving higher reliability estimates for the internal consistency and parallel forms-immediate procedures. The four basic methods of estimating reliability coefficients are discriminated by the time dimension, type of test, and number of testings. The above alternatives only confound time and type of test, checking to see if the student has connected time of testing in both preceding structures to generate the transfer structure relating estimation methods and type of factors.

Comparison of Present Method with Traditional Test Construction

A cognitive structure test differs from the usual achievement test in both the construction process and the final product. Before constructing a typical achievement test the writer usually lists the topics which should be covered with their appropriate weightings. By contrast, an outline of a cognitive structure test is based on the structural analysis of the material, which is therefore dependent upon the elements and relationships identified. Topics and elements are different, although the two domains do overlap. Not all topics would be considered structural elements and not all elements would necessarily be topics. Topics which are discussed but not included in a structural analysis would not be structural elements. It is also possible that the test writer might omit some topics which are structural elements because of personal preference (subjectivity is not necessarily avoided with a structural analysis either), some topics lending themselves very easily to questions, section headings in the text itself (which are not necessarily reflective of structure), etc. Weighting of topics with the usual achievement test is determined by some arbitrary criterion. Weighting of elements and relationships (structures) on a cognitive structure test is determined by the structural analysis of the material, the weighting depending primarily on the

interaction with other substructures. In other words, a structure which is crucial to understanding the material, is at the "heart" of the material, will probably interact with other structures more often than one which is not as central. It will thus be given more weight because the test is constructed to systematically examine all the interrelationships between elements.

When we consider the actual construction of the tests, different processes are also involved. In the multiple choice achievement format, the items have been written to cover important topics, but the crucial problem (for reliability and validity purposes) is to construct good alternative distractors, such that item difficulty and item discrimination indices over appropriate ranges. But these statistical criteria tend to avoid the issue of what type of content, not format, actually makes one alternative better than another. It might be that good distractors are those which contain "confusable" relationships of elements, ones that would be easily detected by a structural analysis of material. But this is only a hypothesis and not one that has been supported by empirical evidence. As stated previously, items on a cognitive structure test are written to test for relationships. But these confusable relationships are not necessarily included to mislead the student, but rather to adequately determine how he has connected ideas within his own cognitive structure. If a student contradicts himself, he obviously has learned and retained something different from a student who responds consistently.

The scoring procedure for these two types of tests also differs.

On achievement tests the answer made to a given item is scored independently of the answer on any other item and these scores are then added. In a sense this procedure reflects the number of things a student knows. It also leads to the possibility that students may get an answer right by chance. Admittedly this chance is not "pure" but the probability of getting an appropriate answer is higher on most achievement tests than on a cognitive structure test. On such a test, the total score reflects the number of correct relationships existing for the student. These scores, as mentioned previously, are determined in part by a pattern of answers, which means that a student is less likely to get an answer totally correct by chance alone, i.e., the probability is less that a student has five items correct than the probability that he has one item correct. Not all items involve patterns of responses even though they are scored for the number of correct relationships, e.g., one item may reflect five relationships and another may reflect two. The total score may also be broken into parts reflecting the basic substructures within the passage.

Traditional test analysis is not directly applicable to this type of test because of the differential weighting given to each item and the items which are scored by patterns. Thus, in general, the usual item difficulty and discrimination indices can not be used. Where sequential dependencies exist between substructures as shown by the structural analysis, answers should reveal a Guttman scale pattern, one set of relationships being a prerequisite

for another. This does not necessarily imply a 50-50 split on difficulty with such items. Rather one would look for patterns within students, expecting either correct-incorrect (+-), correct-correct (++), or incorrect - incorrect (--) sequences, not incorrect-correct (-+) ones. Thus the emphasis would be on patterns which are characteristic of a group of students or perhaps the ones which are not characteristic. The difficulty of such items might range from 0 to 100% and still be consistent with a Guttman scale pattern. Item discrimination indices would not apply to individual items which are scored in terms of patterns. The subscore for a pattern could be used instead, but the weightings of such items with the total score would be greater than the usual achievement test items.

An implicit assumption has been that most achievement tests do not test for structural relationships. Many writers may have such an analysis in mind when they write questions but they probably do not apply it in a systematic fashion. In other, perhaps extreme, cases the structure of the material may not be central at all. On such a test a student may respond correctly based on partial knowledge of the structure or on rote memorization. It is not claimed here that a test based on a structural analysis avoids all problems of test construction, but it does systematically and thoroughly test for knowledge of structure, which could be called mastery. Achievement tests constructed in the usual manner examine only a sample of all possible structural relationships. Thus, depending upon the sample, an individual

could do well on the test without knowing large blocks of material .

However , a test based upon a structural analysis examines all relationships .

In order to perform well , a student must have grasped correctly all the interrelationships between ideas , and probably done some additional relating himself ; partial comprehension will not suffice . Thus it could be assumed that unless a special "set" is given to emphasize structure , ordinary reading would not necessarily result in good performance on such a test although good performance might be expected on most achievement tests .

Comparison of This Approach with Other Approaches to Representing and Testing Cognitive Structure

Neo-behaviorist

This approach to cognitive structure differs from the neo-behaviorist's complex network of S-R associations primarily in that the neo-behaviorists are representing a process and the diagrammatic approach is representing a product . In the neo-behaviorist approach there is a stochastic chain representation of behavior where a given stimulus (covert or overt) leads to a certain response (covert or overt) with a certain degree of probability . With this model the neo-behaviorists give explanations of such behavioral sequences as problem solving , concept attainment , and creativity . The diagrammatic approach to modeling cognitive structure is not a model for such processes . It represents their outcome and does not assume that any one approach best

explains the processes themselves.

Assuming then that the basic difference between the two approaches is process versus product there still remains the question of whether the product is essentially one of S-R associations. A comparison of the basic elements of each approach will answer this question. With the neo-behaviorists the basic unit is "S-R" where "S" represents "stimulus" (covert or overt) and "R" represents "response" (covert or overt), and "-" represents the fact that they are connected with a certain degree of probability so that when S appears R tends to follow, unidirectionality implied. With the diagrammatic approach the basic unit is "O/#" where "O" and "#" represent cognitive elements and "/" represents the type of relationship that exists between them. "O" and "#" do not represent respectively a stimulus and a response, and there is no unidirectionality implied. The symbolization could be "#*" with a different relationship specified. There is no beginning and end implied with this conception of cognitive elements. "/" represents a specific relationship that holds between "O" and "#", this relationship being a learned connection based on the subject matter area from which the elements were also learned. It does not represent a stochastic situation but rather that within a given subject matter two ideas are connected in a certain manner. When applying the diagrammatic approach to testing for cognitive structure, the test checks if the student does have the correct relationship between these ideas. Thus the diagrammatic approach displays a huge, complicated network of

associations, but these associations are not between stimuli and responses, nor is the basic association that of sequential dependency.

In Berlyne's (1965) extension of the neo-behaviorist position, his distinction between structures of a system of symbolic responses as being either bare or stochastic is similar to the previous distinction between process and product.

Bare structure is defined by specifying the responses that are associated as alternative next steps, with each represented stimulus situation at which a subject might arrive. In the case of a transformational hierarchy, this means specifying the alternative transformational responses that can branch out from each situational thought and specifying the new situational thought to which each would lead. Stochastic structure is defined by specifying the bare structure together with the probability of each response that can lead away from a given represented situation (p. 303).

Berlyne stated that structures could be represented by graphs; a bare structure represented by a nonevaluated graph consisting simply of a set of nodes and branches (points and lines).

Bare structure is similar to cognitive structure in that cognitive structure implies specifying all of the possible relationships between cognitive elements that an individual possesses. After testing an individual's cognitive structure in a certain area, only part of the possible set of interconnections has been examined. Some relationships to cognitive elements might be stronger than the ones examined, which would be indicated by Berlyne's bare structure, but the present emphasis is only with testing to see if certain

parts of an entire cognitive structure or bare structure actually exist for an individual .

Berlyne also referred to certain types of structures being quite important in directed thinking, e.g., transitive structures and habit family hierarchies. These structures imply representation by certain types of graphs, completely connected and tree respectively, which is different from the conception of cognitive structure presented here. The present emphasis is not upon what types of relationships are important in directed thinking, but rather upon what types of relationships have been learned from a passage.

Piaget and Reitman

Piaget (Flavell, 1963, p. 17-19, 164-236) has postulated logico-mathematical structures as models of cognitive structures. Cognitive structures refer, in this case, to the organizing properties of intelligence. They are created through assimilation of and accommodation to the environment and are inferable from the behavioral content or data which they determine. For Piaget these structures vary with age, and he has represented developmental stages by various logico-mathematical structures which express the essence of these organizational properties. Thus certain cognitive structures are implied from behavior.

If Piaget says that the classificatory behavior of the eight-year old indicates that he possesses the 'grouping of logical classification,' he means that the child's thought organization in the classificatory area has formal properties

(reversibility, associativity, composition, tautology, etc.) very much like those which define this logico-algebraic structure. The latter has certain specific and definable properties; we infer from his behavior that the child's cognitive structure has similar properties. (Flavell, 1963, p. 169).

Many of these structures can be represented by some type of diagram as Flavell has illustrated (p. 180-195).

The present approach differs from Piaget although it is similar in that graphics are used to model or represent structure. Piaget is concerned with modeling the broad area of intelligence rather than certain subject matter areas. It is not postulated here that one type of structure, such as the class of logico-mathematical ones, completely model a given area, but that several types of structural diagrams could be used. The present development does not pertain to developmental theory nor does it postulate that cognitive structure is characterized by structural properties such as reversibility. Such a classificatory scheme might result from the study but it would not be derived from a theory of logico-mathematical structures.

Although Reitman's (1963) primary emphasis was that of simulating thought processes, he did use diagrams to represent cognitive structures. For example, he included various forms of graphs illustrating family relations. But because Reitman was simulating thought processes with the list processing language IPL-V, he usually represented cognitive structures by list form, indicating the individual's organization of a problem. Reitman's hierarchical lists are related to Ausubel's (1963) assumption that each individual's cognitive

structure is organized hierarchically according to the principle of progressive differentiation, going from broad, inclusive concepts to specific, less inclusive concepts.

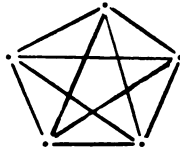
Educational approaches

As stated previously this method of investigating cognitive structure has not been used to any extent in the educational field. Ausubel, as mentioned before, used multiple-choice achievement tests to test for cognitive structure. However, he did not, at least from the reported evidence, construct his tests on the basis of a structural model. Nor did he report in his conclusions whether the varying relationships that subjects had learned might have differed or coincided with the structure of the material.

In Gagné's acquisition of knowledge studies (Gagné, 1963; Gagné and Bassler, 1963; Gagné, et. al, 1962; Gagné and Paradise, 1961) a definite hierarchical pattern of learning prerequisites for a final task was shown, i. e., students exhibited successful learning up to a point and then failed the remaining steps. Yet Gagné's analysis reflected a sequential dependency in learning of topics, not how the topics were structurally related, i. e., no basic relationship between topics was given other than the fact that one was a prerequisite for the other. This is not to say, however, that sequential dependencies do not exist between substructures.

Hartmann (1942, p. 204) graphically distinguished between a

person who knew twenty discrete facts and one who knew five facts in all their permutations and combinations.



versus

However, he did not use this approach in studying cognitive structure.

T. Johnson (1968) reviewed many of the methodological approaches to cognitive structure. His own methodological procedure was based on Ausubel's definition of cognitive structure although the exact correspondence between the two was not clear. Latent partition analysis, a form of factor analysis of categorizations yielding a partitioning of stimulus items into a set of latent categories, was used on concepts from the subject matter areas of teacher behavior and physics. Then a multi-dimensional scaling procedure was applied showing the distances among the latent categories. From these latent categories and inter-category distances inferences were made about how the Ss perceived teacher behavior and physics.

Johnson's approach is similar to the present approach in that he treated cognitive structure as a product not a process. But the procedure did not specify why the stimulus items were grouped together, i.e., the relationships. It was also an attempt to measure how a subject naturally groups concepts (by forcing him into one category sorting, where more than one might be possible) rather than to systematically attempt to test for knowledge relations based upon subject matter. Having individuals sort stimuli into

categories is basic to other procedures attempting to measure cognitive complexity (Scott, 1962; Zajonc, 1960).

P. Johnson (1967) examined relationships between concepts in Newtonian mechanics by using verbal association measure. The formal constraints among these concepts within the subject matter itself was compared with high and low achievers' associations. Again, this procedure only yielded information regarding what concepts were related by a student. It did not reveal how the student perceived them as being connected. Both of the preceding approaches are valuable for instructional guides, but neither provides a way of testing subject matter knowledge.

In this chapter a procedure for testing cognitive structure relationships was explained and compared with traditional achievement test procedures. Transfer structures, covering material that was logically implied by two or more given structures, were also explained. A comparison with other approaches of testing cognitive structure was given. In the next chapter the role of diagrams as perceptual blueprints within learning will be examined.

CHAPTER III

RATIONALE FOR DIAGRAMMATIC PRESENTATION OF STRUCTURE OF KNOWLEDGE

Novak's (1966) model dealt with the transmission of the structure of a subject matter to an individual's cognitive structure and implied that an individual's cognitive structure in some degree approaches the structure of a body of knowledge. Ausubel (1963) and Goss (1961) both emphasized the importance of existing cognitive structure in the learning and retention of additional information. Ausubel (1963) specifically stated that "when we deliberately attempt to influence cognitive structure so as to maximize meaningful learning and retention we come to the heart of the educative process (p. 26)." The main purpose of this study was to examine the effects of different modes of presenting the structure of knowledge upon individuals' cognitive structure immediately after presentation and one week later.

The structure of any area can be represented with language, diagrams or both. Usually language has been used, yet long-term retention and transfer might be enhanced by other modes of presentation. It was expected that the structure of the material would be more clear if diagrams of that structure were presented with the corresponding written exposition of the passage.

The Role of Diagrams in Learning

Sheffield (1961) examined the role of perceptual responses in the learning of sequential tasks. Perception was referred to as a process of interpretation of "filling-in" of sense data, this interpretation depending in general upon past experience. For example, a block of ice which is presented only visually is perceived as cold because in the past it has been sensed cutaneously while being sensed visually. One perceives an airplane just from the distinctive sound made by its motor. A wristwatch is "transparent" to a watch repairman because from the brand and model he can "fill-in" all the internal parts. An important feature of these perceptual responses is that "they permit complete representation of a distinctive stimulus object, even though all of the various stimulus aspects of the object may never be sensed simultaneously (Sheffield, 1961, p. 16)." Other characteristics of perceptual responses are (a) a complete perceptual response can be elicited by a conditioned stimulus in the absence of any of the stimulus aspects of the perceived object and (b) a perceptual pattern can serve as a stimulus as well as a response.

Sheffield specified a type of perceptual mediation which did not require acquisition of a sequence of responses and stimuli but rather that of a pattern. This he called "perceptual blueprinting." Essentially an individual acquires and stores a pattern or blueprint. Behavior is then matched to this memory image or perceptual blueprint. Just as an architect refers back to his

blueprint of a house many times as he attempts to match his perception of his overt product with his perception of this blueprint. an individual uses a memory image in a similar way. It functions as a "blueprint," i.e., the learner manipulates his behavior until his product matches this perceptual blueprint stored in memory. Thus a perceptual blueprint may cue off a sequence of responses. It is important to note that an actual blueprint has advantages over one in memory because the former is exact and unchanging.

Sheffield also stated some other relevant hypotheses.

(1) Most adults are able in some degree to match their overt behavior with their perceptual memories. As such a memory image or perceptual blueprint can serve as a static complex which may be referred to constantly in guiding overt sequences of responses.

(2) There is less interference among perceptual responses than among overt responses. The explanation for this was really not given except that it was assumed that perceptual units are static, rather stable patterns and can be reinstated as fairly complete units without interference from separate parts.

(3) Symbolic behavior can reinstate a perceptual response or unit.

(4) Perceptual units can organize sequences, especially if there is a natural or inherent rather than an imposed organization to a given task. As such it can simplify or aid memory.

The basic assumption in the study was that diagrams in written material function as actual blueprints. When a person stores this pattern in memory, he provides himself with what might be called a memory image or perceptual blueprint and uses this to guide future behavior.

Assuming then that diagrams function as blueprints, they would provide a static blueprint which would guide organization of material during acquisition. Smith and Smith's (1966) review of the role of non-verbal text-book design also supported the position that non-verbal presentations "provide a stable, spatially organized visual framework or background for the more highly articulated and more temporally organized verbal presentation (p.331)" Although diagrams are not entirely non-verbal they do present information in a spatial form. Sheffield (1961) also stated that perceptual units were more effective as organizers of material when there was a natural rather than an imposed organization. Diagrams do represent the "natural" organization of the material and therefore should enhance learning.

From Sheffield's position it can be assumed that when an individual perceives a diagram while reading a passage and the diagram follows explanatory material, the perceptual response will trigger a mediation sequence about the material directly related to it. If the diagram precedes the related material and if it introduces new, unfamiliar terms, the perceptual response might not cue off any relevant mediation. Thus in the first case the diagram would serve as a review or integrator; in the second, only as an introduction to new terms and would not evoke any relevant content.

The organizational advantages of diagrams during acquisition would also apply to memory. Sheffield also hypothesized that there was less interference among perceptual responses than overt responses in memory. This, of course, did not imply that no interference occurs, but it was expected that perceptual blueprints or memory images of the diagrams would be retained with little interference. Thus the diagrams would facilitate retention to a greater extent than only verbal presentation of the material.

When an individual is tested for recall of knowledge through an achievement test, a test item (verbal stimulus) would reinstate the relevant memory image or perceptual blueprint. This in turn would act as a stimulus, in pattern form, to evoke the relevant sequence of mediators necessary to answer the question. Thus diagrams would act as a blueprint, guiding the student's responses and serving as a fairly stable referent. A test can be viewed as a problem of information retrieval. Since perceptual blueprints function as organizational aids in memory and are subject to less interference than other types of memory traces, they would provide a means for retrieving more information. But if the material had not been correctly acquired and distortion had occurred, they would serve as aids for retrieving incorrect information.

Many of these ideas were only tentative. Sheffield provided some support for his position in a study which involved learning two different mechanical assembly tasks (Sheffield, Margolius and Hoehn, 1961). The material was presented by film which utilized a form of perceptual blueprints.

The tasks were divided into sub-assembly units. At the end of each sub-assembly presentation a series of "stills" displaying the parts of the sub-assembly were presented in rapid succession (ikonic level of representation). Each part "jumped" into its proper place in the proper sequence. This technique was an attempt to provide a static blueprint of the material and to guide performance by perceptual memory on a criterion test of assembling the motor. It yielded higher performance scores than a presentation omitting the blueprints. However, this interpretation was not completely clear because repetition effects were uncontrolled.

Another aspect of diagrams, not specified by Sheffield, is that they separate relevant and irrelevant information. Morrison (in Bruner, 1966a, p. 263) viewed translating the sentence "the wind is blowing from the east at 30 miles an hour, " into a diagram (an arrow, ikonic mode) as useful because "it is a noiseless version of the original statement, containing all the information and only the information relevant to the problem (p. 264)."

The role of relevant and irrelevant information has been studied in concept attainment studies. In summarizing these factors Archer (1966) concluded that (a) increasing the amount of irrelevant information decreases the speed of concept identification, (b) inclusion of redundant relevant information offsets the effects of large amounts of irrelevant information, and (c) concept identification will be facilitated when relevant information is obvious and irrelevant information is minimized. The ultimate condition is

where irrelevant information does not exist. Although reading material to attain concepts is different from the usual, more refined concept identification task, it was assumed that the effects of irrelevant information in written material would be similar. Diagrams, as representations of structure emphasizing important relationships and concepts, focus on the relevant information and indicate what is necessary for comprehension of that particular structure. As such they would enhance the learning process.

In summary, it was assumed that during acquisition diagrams would (a) evoke related content if presented after the material necessary for its comprehension, (b) introduce new terms if presented before explanatory material, (c) organize the material in a static, perceptual form and (d) separate relevant from irrelevant information. In memory they would provide rather stable resilient memory traces and serve an organizational function as well. During recall they would act as a readily available stimulus guiding the retrieval of information. With these advantages it was hypothesized that students exposed to diagrams would acquire the structure of the material better than students not exposed to diagrams. Students exposed to diagrams would also perform better on transfer to new tasks based upon the structure of the material. It was expected that the superiority of the diagram presentation would occur on both immediate and later recall.

Individual Differences in Understanding Diagrams

All the preceding major assertions have assumed that the student is able to understand and relate diagrams to the written passage. Vernon (1952, 1953a, 1953b) showed that charts and graphs, which illustrated quantitative data, and pictures were not understood or related appropriately to written text unless subjects, including adults, had training. Graphs were best understood when they related directly to the text. Malter (1948) also concluded that training was necessary and that labels should be used to explain the meaning of graphic symbols such as arrows and dashes. Due to the unfamiliar notation and symbolic nature of the diagrams used in the present study a training program on interpretation of diagrams was constructed.

Of course, individual differences for thinking in spatial forms and retaining such information will cause variations on criterion tests even though individuals have been trained to the same criterion in the use of diagrams. A mental image has been defined as "a more or less complete representation of the attributes of an object or event once experienced but not now present to the senses, together with recognition of its pastness (English and English, 1958)." That people differ greatly in their ability to form mental images is well accepted (Lovell, 1964, p. 97; McKellar, 1957, p. 19), although for most people visual imagery seems to be clearer and to arise more frequently than other types (Lovell, 1964, p. 97). Sheffield

(1961) assumed that most individuals could guide their behavior by perceptual memories. Visual images would be similar to perceptual memories but images would involve creation as well as recall.

Assuming that diagrams were understood, the students possessing a capacity for visual imagery should retain the material longer and perhaps learn it easier than students not possessing visual imagery. Tests for visual imagery are not very reliable nor valid (Woodworth, 1938). Lovell (1964, p. 98-99) stated that visual imagery is clearly connected with mathematical ability in some way but the form of the relationship is unknown. Perhaps students accustomed to symbolic notations, e.g., mathematicians, would be at an advantage with diagrams. Thus although visual imagery might aid individuals in using diagrams, it was not possible to find appropriate measuring instruments to detect people in this category.

Organization of Diagrams Within a Passage

Because diagrams were to be used in the experimental materials, the organization of these diagrams within a passage was important. It was assumed that the most appropriate placement of diagrams within a passage is determined by both the size and the function of the diagram. Since diagrams are related to the content of written materials, the size of the diagrams is to some extent dependent upon the amount of information they encompass. Large diagrams could represent the structure of an entire passage, and small diagrams

could represent substructures. If a diagram is placed before a passage it functions as an overview; if placed after, it functions as a review.

Sheffield mentioned the advantages of diagrams as reviews.

Ausubel's analysis of reviews and overviews also corresponds to Sheffield's position, although Ausubel considered only verbal modes of presentation of material. For Ausubel (Ausubel and Youssef, 1965) both reviews and overviews achieve their effects partly by repetition. Repetitions of the material serve as consolidators of information, as feedback mechanisms to test correctness of knowledge previously acquired, and as sensitizers to the full meaning of the material. Reviews consolidate information and enhance material which is highly available (Ausubel, 1966), while overviews prefamiliarize the learner with certain key terms (Ausubel, 1963, p. 214). This analysis corresponds to the perceptual response analysis of Sheffield; overviews being relatively ineffective when many new terms are introduced but not yet fully explained, making much of the introduction meaningless for the students. The preceding statement assumes that reviews and overviews are not of the name or list type, e.g., "This chapter will cover (has covered) S, U, and V, " but that they do cite basic concepts and their interrelationships. Diagrams are not necessarily classified as a type of advance organizer (Ausubel, 1963). Advance organizers relate and compare concepts the individual already possesses to the new material thereby activating appropriate, rather stable pre-existing, subsuming concepts in an individual's cognitive structure for the new material. Diagrams

do not purposely integrate the new with the old, however they could be used in this way.

Experimental studies have found advantages for reviews, have revealed little on the utility of overviews (because of poor experimental designs), and have not provided an adequate comparison of the relative effects of reviews and overviews. Christenson and Stordahl (1955) obtained no facilitating effects with reviews or overviews, but the experimental material was short and highly familiar to the subjects and the same test was used for pre and post testing. Reynolds and Glaser (1964) investigated massed and spaced review treatments by repeating technical terms on a specified ratio basis in the context of a condensed version of the original topic. The spaced review yielded higher performance than did massed review, and both types of review were better than the no review group. Merrill and Stolurow (1966) found that a hierarchical review did not take more time than when a review was absent, but did result in higher test scores. These studies used verbal reviews and overviews, not diagrammatic ones, but the repetition functions should be similar for diagram forms.

From another viewpoint individuals possess a limited capacity for processing information (Miller, 1956; Posner, 1965). Placing the entire structure either at only the beginning or only the end of a passage would probably exceed the individual's coding and processing capacity. Since placement of a large diagram before a passage would probably exceed the

individual's processing capacity, introduce unfamiliar terms, not evoke relevant mediation, not consolidate material and since no special advantages for overviews have been found in experimental studies, this procedure was eliminated from the experimental materials in the present study.

Having only one large review diagram was avoided in the present study because the sudden introduction of several complex, inter-connected diagrams would prevent adequate coding even though the terms would be familiar. Placing small diagrams within the material itself would facilitate coding, but would not provide a picture of the total structure of the material. Therefore small diagrams representing substructures were placed within the material and a large review diagram comprised of these substructures was placed at the end. The small diagrams were placed after the corresponding verbal passage thereby functioning as a review. This analysis was supported by one pilot subject who had the small diagrams before the corresponding verbal passage, yet who did not examine the diagrams until the relevant material had been read anyway.

Overall Sequence of a Passage

According to Ausubel (1963) the overall sequence of topics in a passage should be arranged by the principles of progressive differentiation and integrative reconciliation. Newton and Hickey (1965) found that learning was facilitated when subconcepts, used in defining another concept, were

placed with the concept rather than separated from it. Gagné (1963) applied the question "what would the individual have to be able to do in order that he can attain successful performance on this task, provided he is given only instructions?" in order to arrive at his hierarchical sequence of tasks. All of the preceding methods imply a highly organized text, with no detached items, the place of each part determined by a broader scheme, and the hierarchical interrelationship of parts being rather clear. In general, a familiar (often means rather inclusive concepts) to unfamiliar (specific or technical) sequence is implied with prerequisites placed as such. These principles were used as guidelines in construction of the written material. Within this broad sequence, rules or generalizations were presented first followed by relevant illustrations and examples of them.

In this chapter the role of diagrams as perceptual blueprints in learning, memory, and recall was presented. The appropriate placement of diagrams within passages of material was also discussed. In the next chapter the results of the pilot study will be described, and the hypotheses for the main experiment will be given.

CHAPTER IV

PILOT STUDY , EXPERIMENTAL MATERIALS , AND HYPOTHESES

First the pilot study will be outlined, then the experimental materials will be described including pilot study revisions, and finally the hypotheses will be given .

Pilot Study

Fourteen Ss (6M, 8F) , volunteer students from a senior undergraduate learning course at Michigan State University , participated in the pilot study . Although the sample for the main experiment was drawn from a senior undergraduate learning course at the University of Alberta, the differences between these two samples were assumed to be minimal .

The Ss were assigned to three experimental treatments, differing in mode of presentation of the structure of a passage on reliability of measurement. . One presentation mode involved diagrams of the structure (D), one had verbal statements instead of these diagrams (V) , and the other was a passage without diagrams or verbal statements, i.e. , no-review (NR) . A complete description of these materials is given later . Four dependent variables were examined: the existence of cognitive structure relationships,

transfer based on these relationships, achievement measured by a typical achievement test, and material incidental to the structure.

Most Ss (8; 4M, 4F) received the diagram treatment because it was expected that Ss would have difficulties with this new mode of presentation. Both sexes were distributed as evenly as possible across both the experimental treatments and the progressive revisions of the materials. Those Ss who received the diagram version of the reliability passage also had the diagram interpretation training program.

The pilot study was used to clarify instructions; to revise the diagram interpretation training program, the three versions of the reliability passage, and the four tests; to give time estimates for the materials; and to determine placement of a formal break within the materials. Ss' answers to questions (Appendix A) and their spontaneous comments about each of the materials were used to produce several revisions. The experiment was explained to all Ss after participating.

Experimental Materials

Diagram Interpretation Program

Goss (1961) mentioned that presentation modes could often be acquired somewhat independently of the content itself. For example, the lines partitioning a Venn diagram represent subdivisions of a larger class and this can be learned somewhat independently of specific content. Such

principles of representation, as well as the relationship between verbal material and graphic representations of it, were emphasized in the training program (Appendix C, final form). It included the three types of diagrams used in the reliability passage: Venn diagrams, tables or matrices, and graphs. Both elements and relationships were included although more emphasis was placed on relationships. In some cases Ss were asked to interpret diagrams without the corresponding verbal material being presented.

A branching programmed learning format was used for the training with questions given after examples of the various diagrammatic forms. Questions always presented a two-choice situation, one choice correct and one choice incorrect. Students were instructed to mark their answers and then turn to a specific page, which in turn presented feedback and elaboration on their answer. If they answered incorrectly they were not instructed to answer again but were given corrective information and then instructed to proceed to another page. The content was different from the reliability passage itself. Since Ss marked their answers to each question a record of the number of errors for each student was available.

The interpretation training program required two major revisions. Some Ss had difficulty interpreting the small diagrams within the reliability passage which represented systematic factors and constant and varying unsystematic factors. These diagrams involved overlapping of classifications and interpretation of a pattern of checks. To facilitate this interpretation a

section on overlap (intersection) of classes was included in the Venn diagram section of the training program. In addition a passage on the importance of a pattern of checks was inserted in the chart section of the training program.

Reliability Passage

The passage was a rather non-statistical, yet technical treatment of reliability based on Ghiselli (1964). The only statistical concept presented was that of correlation, introduced through scatter diagrams. No formulas or calculations were given. An outline of the passage is given in Appendix D. The material was not as highly structured as areas such as chemistry or mathematics, yet Ghiselli's treatment of types of score variation yielded nicely to a structural analysis. The topic of reliability was chosen because it was an area which the experimenter had studied in some depth and was at an appropriate difficulty level for the Ss. The passage dealt with the importance of reliability and of types of score variation, developed a definition of reliability, and examined practical ways of estimating reliability coefficients.

The 7,000 word passage, without diagrams or verbal statements, is presented in Appendix E (final form). Placement of diagrams or verbal statements is indicated in the right hand margin. In the diagram version 19 small diagrams were presented after the paragraphs explaining the subject matter structure they represented. The maximum number of relationships in

any small diagram was seven and the minimum was two . At the end was a large review diagram which combined these small diagrams into six sub-structures and also inter-related these substructures . The diagrams are presented in Appendix F in the same order as they appeared within the passage . The small diagrams within the passage had no explanations because the training program and reliability text provided sufficient guidance .

In the verbal statement condition , the diagrams were replaced with short , separate , single-spaced paragraphs stating very concisely the material of the diagram itself . These passages are also presented in Appendix F following the corresponding diagram . The same labels were used in the verbal statements and diagrams so the criterion test was not biased in favor of either . Both the verbal and diagram reviews were arranged in the same order which did not deviate greatly from the reliability passage sequence .

The major revisions in the reliability passage resulting from the pilot work involved rewriting the sections on systematic factors and on constant and varying unsystematic factors; providing additional directions on reading the final review diagram and some diagrams within the passage; and including six additional major sub-diagrams and then placing these sub-diagrams within the reliability passage at appropriate review places . The sections on different types of factors gave more examples of each type , specified more clearly their similarities and differences and provided additional summary statements . The six sub-diagrams in the final review diagram were numbered and a

suggested order or proceeding through the diagrams was given .

In the first version of the diagram passage , 19 small diagrams were presented within the passage and then combined into six sub-diagrams for the final review . Since Ss had difficulty interpreting this final review diagram because many combinations of the small diagrams had been made simultaneously , the six sub-diagrams (marked "Sub-D" in Appendix F) were placed at appropriate review points within the passage . For example , after small diagrams relating to (a) types of variation in scores , (b) score arrangement and types of variation , (c) relation of factors and variation , and (d) types of unsystematic factors had been presented , the sub-diagram which combined these four smaller diagrams was presented and integrated into the text in the following way: "At this point we can now briefly re-examine the types of factors and their respective outcomes with the diagram given below ." Thus Ss saw the six sub-diagrams within the passage before they encountered them in the final review . Corresponding changes were made in the verbal passage to equate for repetition effects . With this change the verbal and diagram passages provided four presentations of the structure of the reliability passage: (1) the passage itself , (2) the 19 small verbal or diagram sections , (3) the six larger verbal or diagram review sections , and (4) the final verbal or diagram review . The straight version had no summary or repetition of the structure .

Tests

Only one test (Appendix G, final form) was administered, it being originally divided into four parts measuring cognitive structure relationships, transfer based on those relationships, incidental material, and achievement.

Cognitive Structure Relationships

As stated before only cognitive structure relationships, not elements, were tested. Elements were usually given in the item stem, the correct alternative indicated the relationship involved. In order to test systematically for all relationships, the items were predominantly multiple true-false, i.e., a multiple choice format where each alternative was marked true or false. In some instances when a student gave incorrect answers it was possible to determine exactly how he had incorrectly related two elements. In other cases the testing procedure was not complete enough to determine exactly what relationships did exist, although the range could be narrowed to several possibilities. The scoring was based on the number of correct relations; one item yielded as many as six correct relations while in another situation as many as five items yielded only one. The items were based upon the six main substructures that constituted the review.

In some cases patterns of responses were scored. As mentioned

previously the patterns could be classified as consistent and inconsistent . Inconsistent patterns and consistent , but wrong , patterns imply different things about cognitive structure . In the first case the subject has contradicted himself , while in the second case the student has learned something well with no contradictions but learned it wrong . In order to distinguish between these two situations , inconsistent patterns were scored minus one point ; consistent but wrong patterns , zero points . This scoring procedure was used mainly to distinguish inconsistent from consistent but wrong patterns and not to indicate that one pattern was necessarily worse (from a learning viewpoint) than another . It was assumed that the one point difference between the two scores would not greatly affect the comparison among the total scores for individuals . The other consistent patterns were scores for the number of correct relationships they reflected . A complete analysis of each item is given in Appendix G . This analysis includes reference , by diagram , to the substructure being tested , the analysis of response patterns , and the score for each item .

Transfer

Transfer items tested new material which had as its prerequisite the availability of the correct relationships between elements presented in the reliability passage . The transfer topics were not mentioned in the reliability passage but required a thorough understanding of its structure to be

answered correctly . These items were scored similarly and are also presented in Appendix G with a complete analysis of each . It was hoped that this type of transfer task would distinguish between those who had comprehended and retained the material and were able to use it as a basis for new information and those who had comprehended and retained the material but could use it only in the restricted limits of the text itself .

Achievement and Incidental

Ordinary achievement items were included to provide a base-line . These items were constructed from the no-review version of the reliability passage by a doctoral student majoring in tests and measurements . These items are included at the end of Appendix G . A few items covering incidental material , topics not relevant to the structural analysis of the reliability passage , were originally included but eliminated after the pilot study because they served a function similar to the achievement items .

All items (achievement , structure , and transfer) were ordered on the test so that previous ones did not provide answers to later ones . However , it was difficult to entirely eliminate overlap of questions since most of the test was quite comprehensive . The order of the items and maximum score for each test are given in Appendix H . Some minor test item revisions were found to be necessary from the pilot work . These consisted mainly of instructions for answering the items , grammatical changes , and clarification of item stems and alternatives .

Additional Pilot Study Results

The interpretation training program required an average of 28.4 minutes (range 25–31 minutes, $n=8$); the reliability passage, an average of 47.57 minutes (range 35–70 minutes, $n=14$); and the test an average of 50.21 minutes (range 35–75 minutes, $n=14$), yielding an estimate of two hours for the entire experiment. Because of the small sample size, time estimates for each of the three treatments were highly unreliable and not generalized to the final sample.

The formal break was placed preceding the section on constant and varying unsystematic factors (see Appendix D). This section, in the middle of the reliability passage, appeared to be quite difficult for most subjects. With the combined effect of increased difficulty and fatigue Ss tended to skip the section or not attend to it as well as on preceding sections. It was expected that a break would increase attention and decrease fatigue.

Any conclusions about differences between the experimental treatments from the pilot study was not warranted because of the small sample size and the progressive revisions that were made of the materials. However, on the final revisions of the materials the Ss in the Diagram treatment were performing at a higher level than the Ss in the Verbal treatment. The means and ranges for these two groups of Ss on the major dependent variables were as follows (Diagram treatment presented first): Structure; 78 (range 70–86), 68.5 (range 63–74), Transfer; 37 (no range), 23 (range 21–25), and

Achievement; 23.5 (range 22–25), 20.5 (range 20–21). After the pilot study was completed a questionnaire (Appendix B) was designed for the main experiment asking how Ss used the diagrams or verbal statements, whether they liked the passage, and what parts of the passage were difficult. These parts referred to the six basic substructures which had been identified by the structural analysis.

Hypotheses

The central hypothesis was that diagrams would facilitate acquisition of the structure of the material and transfer to new material following from that structure. It was assumed that diagrams would serve as perceptual blueprints, separating relevant from irrelevant aspects more clearly than verbal statements, organizing the material during acquisition and retention, representing material in a rather stable form for storage, and aiding retrieval of information. The verbal passage controlled for repetition effects of the diagrams and was expected to have some advantage over the straight presentation of the material.

It was expected that diagram presentation would enhance both acquisition and retention of cognitive structure relationships compared to the verbal and no-review presentations. Since transfer required knowledge of structural relations, it was expected that the diagram group would also be superior on this variable of both testing periods. However, an interaction

effect between the experimental conditions and testing time was expected for these two variables with the diagram condition resulting in a smaller retention drop over time than the verbal and straight conditions. Performance under all three treatments was expected to drop with time.

Differences among the experimental conditions were not expected for the achievement variable, since analysis of the achievement test showed that complete knowledge of structural relationships was not crucial for good performance. No differential drop in performance over time among the three experimental conditions was expected on this variable although all would drop.

Substructures

Several Guttman scale patterns were identified among the substructures, and it was hypothesized that Ss' response patterns would also exhibit the same dependencies. In general these took the form of "A + B + . . . I" were prerequisites for correct performance on "Z". These Guttman dependencies (Appendix I) were:

- A. Questions 4 and 5 jointly dependent upon question 6.
- B. Question 7 (second part) dependent upon questions 6 and 7 (first part).
- C. Questions 8,9,11 each dependent upon questions 6 and 7-first (7-second) and 15a and/or 19d.

All of these items centered on the various types of systematic and unsystematic variation. Questions 4 and 5 were application items; the rest were transfer items. It was expected that a higher proportion of Ss (of those who successfully completed the prerequisite items) in the diagram group would perform successfully on these criterion tasks than Ss in the other treatments.

Several questions were exploratory in nature. As stated before six basic substructures constituted the "elements" of the total structure of the material:

- Sb1 Degree of reliability and its practical importance
- Sb2 Definition of systematic factors and variation and of unsystematic factors and variation
- Sb3 Effects of all types of factors
- Sb4 Definition of reliability, reliability coefficient, and correlation coefficient
- Sb5 Parallel forms versus parallel tests
- Sb6 Methods of estimating reliability coefficients

There was no reason to believe that the diagram group would be superior on all substructures on acquisition or retention, but exactly which substructures would be crucial was not hypothesized. Similarly the retention drop on each substructure was of interest.

Expected Correlations

Additional information about the relationships among the various variables was necessary, although it was difficult to specify in much detail the patterns that might appear. All variables were classified into three groups: background, experimental, and dependent.

Background

Sex
Age
General Aptitude
 Quantitative Ability
 Verbal Ability
University Major
Tests and Measurements Background

Experimental

Errors on Training Program
Time on Reliability Passage
Questionnaire

Dependent (Acquisition and Retention)

Cognitive Structure Relationships
 Six Substructures
Transfer
Achievement

The expected relationships among these variables are explained below.

First consider the relationship among background variables and experimental variables. Little variation was anticipated on errors on training program so this variable would not be highly related to other variables. If there was a range of time spent on the reliability passage it would probably be correlated with verbal ability. The only relationship between the questionnaire data and background variables was that having a tests and measurements

course would be correlated with familiarity with the material on reliability .

The relationships among background variables and dependent variables were expected to be similar on both acquisition and retention . Verbal ability and general aptitude would probably be related to achievement scores . If quantitative ability or being a math or science major in any way reflects visual imagery they would be related to cognitive structure relationships and transfer for the diagram treatment .

The experimental variables of training errors and passage reading time were expected to be related to performance on the dependent variables only if their ranges were large . For the questionnaire data several relationships were anticipated: correlation between subjective difficulty with substructures and actual difficulty as indicated by test scores , correlation between examination of the review passages and scores on the structure test , and correlation between familiarity with the topic of reliability and achievement , structure , and perhaps transfer .

The intercorrelations among the dependent variables themselves were expected to be positive and to be similar for acquisition and retention periods . Cognitive structure relations and transfer scores would be related . The substructure on definition of types of factors and variation (Sb2) would be related to the substructure on effects of these factors (Sb3) . Because of the scoring procedure there was a built-in dependency between each substructure score and the total structure score , so some relationships would

exist . Since the substructure on the effects of types of factors (Sb3) was related to many of the transfer items a relationship was expected between Sb3 and transfer . As explained previously achievement was not expected to be highly correlated with structure and transfer . High stability coefficient within each dependent variable were expected , e.g. , transfer on acquisition and transfer on retention .

In the next chapter the experimental procedures will be described . This is followed by presentation of the major results .

CHAPTER V

PROCEDURE AND RESULTS

In this chapter the administrative procedures are described , followed by presentation of data relevant to both the major and minor hypotheses of the study . Chapter VI presents a structural analysis of the relationships which the subjects acquired on each substructure and on each transfer item . Discussion of all results is found in Chapter VII .

Procedure

Subjects

A total of 234 Ss (127 F , 107M) , all undergraduates enrolled in a senior learning course at the University of Alberta , participated in the study . This sample consisted of three groups: two groups (O and C) were students from the experimenter's own classes and the third group (R) was from another instructor's class . Groups O and R both (n=156) received the three experimental treatments , while Group C (n=78) served as an additional control group which did not receive any treatment but took the tests on achievement , structure , and transfer .

Groups O and C were formed in the following manner . Students from three of the experimenter's classes , those who could participate in a

three hour evening experimental session, were assigned randomly by sex to the three experimental treatments. One hundred and forty-two Ss comprised this original group but there was an unexpected attrition of 27 Ss at the experimental session itself, leaving a total of 113 Ss who constituted Group O. All of the students who were in the experimenter's classes but did not participate in the experiment yet took the tests comprised Group C.

Due to the unexpected attrition of Ss from the original experimental group and the necessity for large sample sizes for the item analysis, students from another class were asked to participate (Group R, $n=43$). These Ss were also randomly assigned by sex to the experimental treatments. Subjects were assigned so that the number of Ss and the sex distribution of Ss per treatment were balanced for the total experimental sample (Groups O and R). Table 1 gives this distribution. Group R was given the experimental treatment and retention test one week after the corresponding experimental administration for Group O. It was decided to combine the two groups for statistical analysis if they appeared similar on the main dependent variables on both acquisition and retention periods. Otherwise, separate analyses would be required.

Treatments

All Ss in Groups O and R received the diagram interpretation program, followed by one of the three versions of the reliability passage

(Diagram, Verbal, or Non-Review), with the acquisition test at the end. All Ss progressed at their own rate and finished within a three-hour period. The test which measured cognitive structure relationships, transfer, and achievement was used for acquisition and retention. Immediately after the retest one week later Ss answered the questionnaire about the experiment. Group C took the test when Group O was administered the retest.

Table 1
Distribution of Subjects

		Acquisition	Retention
Diagram	O	38(23F, 15M)	35(21F, 14M)
	R	16(6F, 10M)	16(6F, 10M)
	Total	54(29F, 25M)	51(27F, 24M)
Verbal	O	38(20F, 18M)	34(20F, 14M)
	R	13(7F, 6M)	10(5F, 5M)
	Total	51(27F, 24M)	44(25F, 19M)
No-Review	O	37(17F, 20M)	34(15F, 19M)
	R	14(8F, 6M)	13(7F, 6M)
	Total	51(25F, 26M)	47(22F, 25M)
Control		78(46F, 32M)	

The administration of the experimental treatments differed slightly for Groups O and R. For Group O the three treatments were given in separate rooms by different administrators. Because the administrator only handed out the materials, differences among them were assumed to be

negligible . Different rooms were used to avoid contamination effects by the anticipated differences in reading time for the three treatments . One week later Ss were given a surprise retest during the regular class period . However , for Group R the experimental treatments were given in one room (for convenience) and the Ss were accidentally told of the retest .

The written instructions for the experimental materials were the same for all treatments and are given below .

Diagram Interpretation Program

"In much written material diagrams or drawings are presented . In the following passage you will be introduced to various types of diagrams , each of which is directly related to the content in a written passage . It is the purpose of this instructional program to enable you to interpret these various types of diagrams and to relate them to the corresponding written passage .

"This instructional program is presented in a programmed learning format . After an example of a diagram you will be asked at least one question over it . Answer the question , mark your answer on the sheet itself , and then follow the directions which will tell you the number of the next page to read . In this manner you will progress through the entire program . Please follow the instructions carefully ."

Reliability Passage

"You will now be asked to read a passage on the concept of reliability . Read it carefully for comprehension and read it only once . You will be asked to time yourself on this passage .

"On page x (exact page depended upon presentation mode) a "break" is indicated . When you reach this point in the passage , you may take about a 15 minute break in the hall outside this room . Please do not discuss the experiment with anyone during this period . Please indicate times of stopping

and beginning again on page x where the break is indicated .

"After completing this passage you will be given a series of questions over it ."

Test

"You will now be asked a series of questions over the passage on reliability . There is not a constant format for all of the questions; they are not all true-false , nor all multiple-choice . In general the test is objective .

"Since each question differs in the way in which you should answer it , please follow the instructions very carefully for each question . All answers are to be marked on the test itself .

"Please do not discuss the experiment with anyone else who has participated . Thank you very much for participating ."

Major Results

The following major findings are presented: comparison of the experimental treatments on the main dependent variables of time , errors , achievement , structure , and transfer on acquisition and transfer; the sequential dependencies among items; and the intercorrelations among time , errors , achievement , structure , and transfer .

Data for each hypothesis is presented first for Group O and then for Group R . Differences and similarities between the two groups are mentioned . When appropriate , comparisons with the control group , Group C , are made . If a complete presentation of data is not in the text , it is located in an appendix . The following notation will be used for the

experimental treatments and dependent variables:

Groups O and R, No-Review treatment - O-NR and R-NR
 Groups O and R, Verbal treatment - O-V and R-V
 Groups O and R, Diagram treatment - O-D and R-D
 Time - Tm
 Errors - E
 Achievement on acquisition and retention - A1 and A2
 Structure on acquisition and retention - S1 and S2
 Transfer on acquisition and retention - T1 and T2

Comparison of Treatments on Time, Achievement,
 Structure, Transfer, and Errors

Table 2 presents means and standard deviations for the major dependent variables for each treatment. Separate analyses for each variable is presented after this table.

Table 2
 Means and Standard Deviations for the Main
 Dependent Variables for Each Treatment

		O-NR	O-V	O-D	R-NR	R-V	R-D	C
Tm	M	33.22	41.38	46.53	36.36	34.92	40.98	
	SD	5.41	9.98	8.57	8.81	6.07	10.98	
A1	M	15.38	15.53	15.37	16.50	16.31	17.06	11.06
	SD	3.23	3.67	8.81	3.88	3.29	3.36	3.31
A2	M	14.77	14.74	15.23	15.00	15.50	16.86	
	SD	3.89	3.61	3.27	3.82	4.67	3.08	
S1	M	61.32	62.47	62.74	64.93	60.85	64.50	51.81
	SD	11.52	8.95	11.64	7.99	11.35	10.73	12.31
S2	M	60.24	63.38	61.09	66.46	61.70	63.25	
	SD	9.79	8.15	9.78	7.24	10.01	7.83	
T1	M	24.73	25.16	22.61	22.36	23.92	25.38	21.65
	SD	8.33	9.70	7.42	6.60	5.74	8.45	7.44
T2	M	24.03	25.41	23.29	24.07	28.20	26.44	
	SD	7.26	6.68	6.39	8.53	6.78	6.86	
E	M	.62	.66	.34	.29	.92	.38	
	SD	1.11	.97	.53	.82	1.04	.62	

Time . -- Analysis of variance for the time measurements

indicated that the experimental treatments in Group O differed in average time to read the reliability passage. The F ratio of 26.60 (df 2,110) was significant beyond the .001 level and the treatments accounted for 32% of the total time score variance (Appendix J). Scheffé's multiple comparison technique showed that the D and V treatments each required more time than the N-R treatment (at .001 level), but that the D treatment did not differ from the V treatment (Appendix J). Differences in time had been expected because the treatment passages differed in length.

However, differences among treatments on time for Group R did not occur ($F = 1.29$; df 2, 40; Appendix J). Since the order of the Tm means differed by treatment from that of Group O, the two samples were compared. A one-way analysis of variance with the six treatments (two groups, each with three treatments) was calculated. The F ratio was significant, ($F = 10.90$, df 5,150, $p < .001$) and 27% of the time score variance was accounted for by the treatments (Appendix J).

In comparing Group O and R Scheffé's multiple comparison technique showed that the following treatments differed in average time to read the reliability passage: O-D greater than R-V and R-NR, and R-D greater than O-NR (Appendix J). These differences would have been expected if the two samples were similar on time to read the passage. However, three other comparisons should also have been significant if the

samples had been similar (O-V greater than R-NR, R-D greater than O-V and R-V greater than O-NR).

The basic difference between the two groups was that for Group R the NR treatment required more time than expected and the D treatment required less time than expected. There was less variance among the three treatments times for Group R than for Group O. This discrepancy can apparently be explained by the differences in administration procedures for the two groups. A further discussion of this is given later.

Since time spent reading the reliability passage was considered an important factor in learning and the two groups differed on this measure, the samples were separated rather than combined in the remaining analyses. Thus the attempt at increasing the sample size for statistical analyses failed, yielding instead a replication of the study.

Achievement. -- There was no difference in mean A1 scores for the treatments within either Group O or R (F ratios being respectively: $F(O) = .06$; $df\ 2, 110$; $F(R) = .17$, $df\ 2, 40$). Analyses of covariance with Tm as the covariate also indicated no significant differences ($F(O) = .166$, $df\ 2, 109$; $F(R) = .095$, $df\ 2, 39$). Complete data is given in Appendix K. The correlations between Tm and A1 for all treatments were generally low, ranging from $-.397$ to $.456$. The Kuder-Richardson 20 reliability coefficient for the A1 test was $.64$.

However, the means from Group R were generally higher than

those from Group O so a comparison of these groups as well as of the control group was made. Appendix K shows this analysis of variance. The treatments were significantly different ($F = 15.98$, $df\ 6,227$, $p < .001$) and accounted for 29.7% of the total score variance.

Scheffe's multiple comparison test showed that each treatment in the experimental groups (O and R) differed significantly from the control group C. However, none of the comparisons between treatments in O and R were significant. The comparison of these means is given in Appendix K.

These results were as originally expected. However the difference between the treatments and the control group was not as large as expected even though it was significant. On the average the treatment means indicated that the Ss were performing between chance (7.5) and the optimum difficulty level (19).

There were also no significant differences for mean scores on A2 for Groups O and R ($F(O) = .20$, $df\ 2,100$; $F(R) = .88$, $df\ 2,36$; Appendix K). The reliability of the A2 test was .56 (Kuder-Richardson 20). Analysis of covariance with A1 as the covariate also indicated no significant differences among the treatments for Groups O and R ($F(O) = .36$, $df\ 2, 99$; $F(R) = .673$, $df\ 2, 35$; Appendix K). The correlation between A1 and A2 ranged from .435 to .927.

When Groups O, R, and C were compared the experimental

groups maintained significantly higher scores than Group C ($F = 11.80$, $df\ 6, 213$; $\eta^2 = 24.6\%$). Thus contrary to expectation, little retention drop occurred on the achievement variable. Appendix K gives the analysis of variance and indicates the significant differences between mean comparisons. All experimental treatment means were significantly different from the Group C mean as with acquisition, although the order of the means differed from acquisition.

Structure. -- There were no differences in mean S1 scores for the three treatments within either Group O or R ($F(O) = .15$, $df\ 2, 110$; $F(R) = .65$, $df\ 2, 40$; Appendix L). Analysis of covariance with Tm as the covariate also indicated no significant differences ($F(O) = .095$, $df\ 2, 109$; $F(R) = .584$, $df\ 2, 39$; Appendix L). The correlations between Tm and S1 ranged from $-.343$ to $.445$.

However, the means for Groups O and R appeared to be higher than the Group C mean, so an additional comparison was made with these groups (Appendix L). The F test ($F = 8.09$, $df\ 6, 227$) was significant at the .001 level with the treatments accounting for 17.6% of the variance. Comparison of the means showed that all except one of the experimental treatments scored higher than the control group.

Thus the results on S1 were not as expected; the D treatment was not superior to the V and NR treatments. In general the experimental treatments were superior to the C group; but these differences were not as large as expected.

There were no significant differences in mean scores on structure retention (S2) for either group O or R ($F(O) = 1.02$, $df\ 2, 100$; $F(R) = .95$, $df\ 2, 36$; Appendix L). Analysis of covariance with S1 as the covariate also indicated no significant differences among the treatments ($F(O) = .125$, $df\ 2, 99$; $F(R) = .102$, $df\ 2, 35$; Appendix L). The correlations between S1 and S2 ranged from .137 to .679. When Groups O, R, and C were compared, the experimental treatments generally maintained significantly higher scores than Group C ($F = 9.00$, $df\ 6, 213$, $p < .001$) over the retention period (Appendix L). The treatments accounted for 20.2% of the total score variance.

Again, the results were not as expected. The experimental treatment means did not differ on retention and little forgetting occurred, as indicated by the general superiority of the treatments in contrast to Group C at retention.

Transfer. -- There were no significant differences in mean scores on transfer acquisition (T1) for either Group O or R ($F(O) = .86$, $df\ 2, 110$; $F(R) = .35$, $df\ 2, 40$; Appendix M). Analysis of covariance using Tm as the covariate, also indicated no significant differences among treatments ($F(O) = .528$, $df\ 2, 109$; $F(R) = .38$, $df\ 2, 39$; Appendix M). The correlations between Tm and T1 ranged from $-.209$ to $.466$.

In contrast to the previous pattern of results, there were no significant differences between the experimental treatments and Group C

on T1, even though the mean score for Group C was lowest. Appendix M gives the analysis of variance on T1 for Groups A, B, and C.

There were no significant differences in mean scores for T2 for either Group O or R ($F(O) = .60$, $df\ 2, 100$; $F(R) = .23$, $df\ 2, 36$; Appendix M). Analysis of covariance with T1 as the covariate also indicated no significant differences ($F(O) = .376$, $df\ 2, 99$; $F(R) = .574$, $df\ 2, 35$; Appendix M). The correlations between T1 and T2 ranged from .047 to .669.

In comparing the experimental treatments with Group C on T2, the F test was significant ($F = 2.40$, $df\ 6, 213$, $p < .05$, Appendix M) but the treatments accounted for only 6.3% of the total score variance. Scheffé's multiple comparison technique did not indicate any pairs of means to be significantly different at the .05 level of significance. Thus the significant contrasts must have existed in a combination of pairs rather than in single comparisons.

The expected differences in transfer among the experimental treatments did not occur. Because the experimental treatments did not differ from the control group, the transfer test appeared to be the most difficult of the three tests. This order of difficulty of tests was congruent with the theoretical formulation, but the experimental treatments were expected to be superior. The T2 means were generally higher than the T1 means, as indicated by the analysis of variance. However the slight rise

did not reveal any differences between pairs of means . Discussions of the unexpected results on the achievement , structure , transfer , and time variables is given later .

Errors. -- As expected there were no differences ($F = 1.41$, $df\ 5, 150$) among the experimental treatments on the number of errors on the daigram training program . Appendix J gives the analysis of variance on errors (E) for the experimental treatments . Although some Ss made three or four errors on the training program , the average number of errors for each experimental treatment was less than one . Thus the hypothesis that Ss would make few errors on the program was supported .

Guttman Dependencies

Several dependencies were expected to exist among the Items on the structure and transfer tests; some Items testing information which served as a prerequisite for successful performance on other Items . Within this collection of Items , one Item (#6) was a prerequisite for all the remaining Items (refer to Appendix I for this analysis) . However , every S failed this Item , eliminating half of the possible patterns . In general the majority of the remaining patterns on the dependencies were -- . Since half of the possible patterns did not exist , an analysis of the remaining patterns would not have been meaningful because an adequate test of the hypothesis was impossible . Therefore such an analysis was not included in the study .

The complete failure on item 6 was not expected. It appeared that something besides knowledge itself was a factor determining performance on the items. Two factors were investigated, the information load of the item and the item format. These two factors are discussed more thoroughly in Chapter VI.

Correlations among Time, Errors, Achievement Structure and Transfer

Correlations among the main dependent variables were originally computed separately for the experimental treatments. However, since very few differences among these correlations occurred across the six treatments, and the previous analyses had indicated that the treatments did not differ on mean scores for each of these variables, the correlational data was pooled across all treatments. Thus, highly stable indices of the relationships among these variables were obtained. Table 3 gives the correlation matrix for these main variables. Since the sample size was so large for these correlations, correlation coefficients which were not high in absolute value were significant (any correlation coefficient greater than approximately .14 was significant at the .05 level). In order to distinguish among correlations which were significant and high and those which were significant and not as high, a rather arbitrary cutting point of .45 was used. Any correlation coefficient of this magnitude indicated that 20% of the variance on one variable could be accounted

for by the other related variable .

Table 3
Correlation Coefficients among the Main Dependent
Variables for all Subjects

	Tm	E	A1	A2	S1	S2	T1
E	-09						
A1	07	-15					
A2	10	-26**	68***				
S1	10	-33***	39***	46***			
S2	11	-23**	29***	46***	54***		
T1	-04	-14	18*	21*	06	11	
T2	-04	-23**	17*	32***	25**	30***	36***

*** p < .001
 ** p < . 01
 * p < . 05

Time spent reading the reliability passage did not correlate with achievement , structure , or transfer on acquisition or retention (Table 3) . Errors on the diagram interpretation program was not expected to correlate with the reliability tests because a small range on the error scores was anticipated . In general the small variability did occur but there were several significant negative correlations (Table 3) . A negative

correlation meant that a high number of errors on the diagram interpretation program was related to a low score on the reliability tests. This relationship is congruent with what one might expect if there were a general ability factor underlying performance on different tasks. However none of these correlations were above the cutting point of .45.

As hypothesized high test-retest correlations existed for achievement, structure, and transfer, with the structure and achievement test-retest correlations being the highest and greater than .45. Contrary to expectation structure correlated more highly with achievement than with transfer. Each of the four correlations between structure and achievement was significant but only two of the four structure-transfer correlations were significant. In fact, achievement and transfer were more consistently related than structure and transfer. Discussion of these relationships is presented in Chapter VII.

Minor Results

Experimental Treatment Comparisons on Substructures

In general few differences among experimental treatments were found on the six substructure scores (Appendix G). On acquisition a one-way analysis of variance with seven treatments was computed. Although the over-all F test was significant at the .05 level for substructures (Sb 1, 2, 3, 4, and 5, Appendix N), Scheffé's multiple comparison

technique indicated only four significant differences between pairs of means. In each case the difference was between an experimental treatment and the control group (For Sb1, O-D was greater than C and O-NR was greater than C; for Sb5, O-D was greater than C and R-D was greater than C).

On retention only the six experimental treatments were compared. There were no significant differences on any of the substructures. The F tests and treatment means and standard deviations for each substructure are given in Appendix N. Thus on acquisition and retention no substructure was learned better by any of the Ss in the three experimental treatments. Any differences that did occur were between the treatments and the control group. The similarity among the treatments was also supported by the structural analysis presented on the next chapter.

No definite predictions were made about Ss performance on the substructures. However, in light of the generally poor performance by Ss on the total structure variable, and the similarity of the experimental treatments on total structure, differences between experimental treatments on substructures were not expected.

Questionnaire Responses

A questionnaire was given to all Ss during the retest period, one week after the initial administration of the reliability passage. Three questions were common to the three experimental treatments. The first asked if the

content of the reliability passage was familiar; the second, if the Ss enjoyed reading the reliability passage; and the third, asked the Ss to check which of the six substructures were difficult to understand. Other questions were given to the Ss in the D and V treatments. The data for these latter questions are discussed below, while the data on the common questions are presented later. Because responses to the questions were similar for both groups O and R within each treatment, the data for both groups was combined.

The Ss in the D treatment answered additional questions on how they had studied the reliability passage. Ss indicated little trouble with the small diagrams within the passage, but about half of them indicated problems with the large review diagram. On the review diagram Ss did not always examine the interconnections between the diagrams in a systematic fashion, in fact, two of them ignored the connections. No one method of using the diagrams while reading the passage predominated; all were used about equally. Only one-fourth of the Ss stated that they visualized diagrams while taking the test, and only one-fourth made any other definite associations between the test and the diagrams. Complete data are given in Appendix O.

The Ss in the V treatment also answered questions pertaining to their mode of studying the reliability passage. Again no method predominated, although repetition was the most frequent usage. Very few, less than one-fourth, of the Ss instantly recognized that a review passage related to a test item. Complete data on the V treatment are also given in Appendix O.

Despite training on diagram interpretation Ss indicated trouble in understanding the diagrams within the reliability passage. This confusion could have possibly been the result of inadequate diagram construction. For example, different sized cells within a matrix could have had different meanings for Ss, double-headed arrows might have been ambiguous, etc. These potential difficulties were pointed out by a S who was an engineer and had had previous experience with problems involved in understanding diagrams.

If Ss had mastered the material they probably would have been more aware of connections between the test items and the structural review passages and diagrams. Interference among the diagrams and passages as well as individual differences in imagery and ways of reading material might also explain some of the questionnaire results. However the above evidence is interpreted mainly as additional support for the hypothesis that the Ss did not comprehend the reliability passage. More extensive treatment of this hypothesis is given in Chapter VII.

Relationships among the Major Dependent, Substructure, Background, and Questionnaire Variables

Originally correlation coefficients were computed separately for the six treatments. However, the correlations across treatments were quite similar. So as with the previous correlational analysis among the major dependent variables, data for the remaining correlations were pooled across all subjects. Explanations of the variables will be presented first.

Two ability measures were available from the students' records: matriculation examination scores and aptitude test scores. Because two aptitude test scores were available, separate analyses were required. The aptitude tests will be discussed later, however, the matriculation examinations will be explained now. Matriculation examinations in subject matter areas are given to high school seniors for university entrance requirements. Since the matriculation exams were revised in 1960, it was not possible to use scores from matriculation tests which were administered prior to this date. Thus for 34% of the subjects matriculation scores were not available. For the remaining Ss common scores were available in English (Eng), Mathematics (Mth), Social Studies (SSt), and Chemistry (Chm).

As mentioned previously several questionnaire items were common to all Ss. These pertained to the familiarity (Fa) of the passage, the attitude (Att) toward the passage, and subjective estimates of the difficulty of each of the six substructures (ED-#). The Ss difficulty estimates on each substructure showed a rather consistent hierarchy (ranking) across both treatments and groups (Table 4). The rank orders were obtained from the number of individuals who checked the substructure as being difficult (Ss were allowed to check more than one substructure). Spearman rank order correlations between Group O and R for each treatment were as follows: for D, $r = .656$; for V, $r = .90$ significant at the .05 level; and for NR, $r = .972$ significant at the .01 level. Kendall coefficient of concordance among all treatments was .569 (significant

at the .01 level). Therefore within each experimental treatment the two groups generally agreed on the difficulty of the substructures and across treatments these orderings were also consistent. In all cases Sb1 was rated as the easiest and Sb3 was rated as the most difficult.

Table 4
Subjects' Rankings* of Substructure Difficulty by
Treatment and Group

	Diagram		Verbal		No-Review	
	O	R	O	R	O	R
Sb1	1	1	1	1	1	1
Sb2	4	5	4	5	4	4.5
Sb3	6	6	6	6	6	6
Sb4	3	4	3	3.5	2	2.5
Sb5	5	2	5	3.5	5	4.5
Sb6	2	3	2	2	3	2.5

* Ranking of 1 represents easiest substructure.

Ranking of 6 represents most difficult substructure.

The other variables which were correlated were sex, enrollment in a measurement course (MC), age, the major dependent variable scores, and scores on the six substructures. Table 5 presents the correlation matrix for all of these variables.

Substructures

It was expected that the six substructures would correlate with the total structure score since each substructure score was part of the total score.

Table 5

Correlation Coefficients among Main Dependent, Substructure, Background, and Questionnaire Variables for all Subjects

	Sb1	Sb2	Sb3	Sb4	Sb5	Sb6	Sb1*	Sb2*	Sb3*	Sb4*	Sb5*	Sb6*	Sex	MC	Age	Enq	Mth	Sst	Chm	Fa	Att	ED-1	ED-2	ED-3	ED-4	ED-5	ED-6
Tm	14	-09	13	04	09	-02	09	02	01	03	09	13	-01	-06	13	01	-02	02	09	-05	-19	07	17	-00	-05	05	04
E	-04	-22b	-32a	-24b	-27a	-05	-14	-13	-25a	-02	-01	-05	10	10	17	-30b	-19	-13	-22c	-02	08	02	08	-05	22b	13	-07
A1	14	29a	29a	21b	37a	05	09	18c	27a	09	13	09	09	-20b	-06	15	21c	33a	28b	27a	-20b	03	05	11	10	14c	03
A2	25b	23a	37a	23b	31a	12	18c	33a	29a	22b	10	11	04	-19c	-06	37a	27b	24c	24c	09	-23b	11	06	15	05	17c	02
S1	56a	57a	55a	60a	36a	44a	32a	22b	34a	22b	16c	28a	-10	-07	-06	18	21c	22c	16	16c	-24b	-07	-01	08	-05	06	06
S2	29a	28a	29a	34a	29a	25b	53a	25b	63a	24b	27b	-10	03	-06	03	31b	13	14	18	17c	-13	-02	-02	13	-02	11	05
T1	-02	12	04	04	13	-06	04	19c	04	-01	21c	-12	12	08	-04	07	15	11	-01	00	-12	-11	03	06	-05	-05	-04
T2	21c	21c	03	14	08	02	20c	23b	11	23b	-00	-14	-00	-06	-09	19	16	08	12	02	00	02	-04	12	-12	25b	-02
Sb1	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Sb2	11	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Sb3	05	19c	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Sb4	05	24b	16c	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Sb5	08	13	26a	10	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Sb6	04	07	21b	26a	13	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Sb1*	30b	11	18c	11	08	16	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Sb2*	15	26b	03	13	08	-01	-01	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Sb3*	-04	18c	54a	16	37a	08	01	13	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Sb4*	13	08	-08	32a	01	16	11	24b	-03	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Sb5*	07	01	11	12	35a	01	-04	09	17c	03	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Sb6*	14	11	15	08	16	34a	03	-03	-01	11	03	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Sex	-04	04	-00	-13	-16c	-14	-05	-01	05	-13	-07	-13	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
MC	-02	-09	-01	-06	-02	-00	-14	11	03	03	14	02	-16c	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Age	10	-12	-05	-07	01	-13	02	05	-01	-17c	05	-03	04	24	..	..	..	..	..	..	..	..	..	..	..	..	..
Enq	00	11	15	14	16	10	16	22c	06	19	07	12	-39a	-09	-25b	..	..	..	..	..	..	..	..	..	..	..	..
Mth	18	21c	-01	13	-02	04	14	13	-21c	10	23c	07	01	-22c	-04	18	..	..	..	..	..	..	..	..	..	..	..
Sst	-09	17	16	26b	13	04	14	18	-09	23c	00	-02	-10	-15	39a	09	..	..	..	..	..	..	..	..	..	..	..
Chm	02	15	07	12	12	12	06	09	04	11	14	19	-07	-03	-08	34a	26c	24c	..	..	..	..	..	..	..	..	..
Fa	-01	03	18c	17c	14	05	09	05	25b	01	11	-03	04	-12	-02	01	-12	13	14	..	..	..	..	..	..	..	..
Att	-18	-13	-23b	-00	-23b	-03	-17	03	-25b	14	-23b	-02	-25b	05	-18	07	-14	-11	-17	-19b	..	..	..	..	..	..	..
ED-1	03	07	-09	-10	-06	-03	-07	08	-05	06	-06	-06	-05	-18c	-07	06	-15	03	-03	-21b	-02	..	..	..	..	..	..
ED-2	05	12	-06	-05	06	-19	06	07	-10	-09	-05	05	-05	-10	14	17	-01	13	06	-00	-12	03	..	..	..	..	..
ED-3	07	12	06	-01	07	-05	06	15	11	00	-01	04	-07	-07	13	36a	14	03	34a	02	-07	-08	22b	..	..	..	..
ED-4	04	-15	07	-07	-10	00	-10	-11	08	05	03	04	18c	-17c	-03	-11	-06	-12	02	10	-16c	08	-13	-19c	..	..	..
ED-5	13	05	-02	-01	-09	02	06	11	-06	16	-00	-07	02	-08	-09	13	01	07	19	10	-12	06	-05	-10	26b	..	..
ED-6	05	-02	08	01	20c	-04	-05	08	15	-05	10	01	-05	01	08	-02	-02	10	09	04	-02	08	11	-09	-09	00	..

* Retention

a p < .001

b p < .01

c p < .05

It was also predicted that Sb 3 would correlate with transfer since the original analysis had shown that this substructure was a pre-requisite for many of the transfer items. Originally it was also hypothesized that the substructure scores would be more highly related to transfer than achievement. However the previous analyses in Chapter V showed that achievement and structure were related and that structure was not related to transfer. Because of the built-in dependency between substructure and the total structure scores, it was clear that the original hypothesis would not be supported. However, the relationship between Sb 3 and transfer might still be high, depending upon the correlational pattern between the other substructures and transfer.

Each substructure correlated significantly with the total structure score on retention and acquisition, as hypothesized. However, the strongest correlations were within time periods, rather than across time periods, i.e., acquisition substructures correlated more highly with acquisition total structure than with retention total structure. The substructures were more highly related to achievement than to transfer, and Sb 3 did not correlate with transfer. The lack of relationship between structure and transfer can be largely attributed to the Ss' inadequate comprehension of the reliability passage and to the format and information load of the items. These factors are discussed in Chapter VII.

Time did not correlate with substructure but errors did, especially on acquisition. Significant negative correlations between errors and substructures meant that Ss who made many errors on the diagram training program

did poorly on the substructures as well. This finding is consistent with the relationship between errors and the main dependent variables.

It was hypothesized that Sb 2 and Sb 3 would be related to each other on acquisition and retention. The relationships among the other substructures were predicted to be low. In general Sb 2 and Sb 3 were not highly correlated (two of the four correlations being significant and none greater than .45). Within each testing period, the inter-correlations among substructures were generally low (8 significant correlations out of 30 and none greater than .45). The lack of relationship between Sb 2 and Sb 3 can be explained by the contaminating factors of format and information load which are discussed in Chapter VII.

It was also hypothesized that correlations between corresponding substructures over time (acquisition and retention periods) would be high. For each substructure significant correlations did occur, supporting the hypothesis. Other correlations between acquisition and retention substructure scores were not significant, as expected. Thus except for the lack of relationship between Sb 2 and Sb 3, the original hypotheses about the relationships among the substructures were supported.

Questionnaire.--The questionnaire items of familiarity, attitude, and subjective difficulty estimates will be discussed in that order.

It was hypothesized that familiarity with the content in the reliability passage would correlate with enrollment in a measurement course, and

would correlate with achievement and structure but not with transfer. However, familiarity and enrollment in a measurement course were not correlated. Some Ss could have been exposed to reliability concepts in other courses, or the content of the reliability passage itself could have been quite different from the treatment of reliability in an undergraduate measurement course. Both of these reasons could apply to the present sample. Some students learn about reliability in learning and individual differences courses. The treatment of reliability in the experimental passage was more technical than is often given in undergraduate courses. These two explanations are somewhat contradictory, but students could have interpreted "familiarity" differently, this making both explanations possible.

Familiarity correlated significantly, but positively, with A1, S1, and S2. A significant positive correlation meant that individuals who were not familiar with the content of the reliability passage scored higher on achievement and structure than those who were familiar with the content. This result was contrary to expectation. Perhaps Ss who thought they were familiar with the content did not study the passage as intensively as those who were unfamiliar with it, thus accounting for the unexpected results. Different interpretations of "familiarity" by Ss could also have confounded the relationships.

It was hypothesized that attitude toward the reliability passage would not be correlated with enrollment in a measurement course and would also not correlate highly with the dependent variables. The first hypothesis

was supported, but the second was not. Significant negative correlations occurred with A1, A2 and S1 and with 5 of 12 substructure correlations. A significant negative correlation meant that those who disliked the reliability passage scored lower on the dependent variables than those who enjoyed the passage. This result is consistent with two general ideas about the relationship between interest and performance: (a) interest increases motivation thus increasing attention while learning and the final performance level and (b) lack of success on learning and assessment tasks decreases perceived interest and motivation.

In general there were no significant correlations among the substructure difficulty estimates. Correlations among the difficulty estimates and the main dependent variables were generally non-significant as well.

Of particular interest was the relationship among estimated and actual difficulty (determined by Ss scores) on the six substructures. It was expected that those who cited a substructure as being difficult would perform poorly and those who understood the content would score high (this relationship being reflected in significant positive point-biserial correlations). However, this was not the situation. Only two of seventy-two correlations were significant for both acquisition and retention (one positive and one negative). One possible reason for the lack of hypothesis support was that the estimation scale (a "yes" or "no") was not sufficiently sensitive to distinguish between the different levels of subjective difficulty that could have

existed for the Ss. Thus the scale could have been broadened. Another possibility would have been to follow-up an initial indication of difficulty with a question which asked the S if he thought he had mastered the material even though he had found it difficult. The small variance on some of the difficulty estimates also attenuated the correlations.

Background Variables.--Sex, enrollment in measurement course, age, matriculation, and aptitude variables will be discussed in that order.

Sex, enrollment in measurement course, and age were not expected to correlate with the dependent variables. The results for sex supported this hypothesis, however two exceptions occurred between enrollment in measurement course and A1 and A2. A significant negative correlation meant that Ss who were enrolled in a measurement course scored higher on achievement than those who were not enrolled. The absolute values of these correlations were not high, but they did indicate that the achievement test was more similar to the content in measurement courses than were the structure or transfer tests. Age correlated with errors on the diagram training program, indicating that the older Ss made more errors than the younger Ss.

Matriculation scores did not correlate with time spent reading the reliability passage, but Eng and Chm scores did correlate negatively with errors on the diagram training program, indicating that those Ss who scored highly on these matriculation exams made few errors. Matriculation scores were significantly related to achievement (except Eng and A1), and there was a

tendency for matriculation scores to be related to structure (three out of eight were significant). However, matriculation scores were not related to transfer.

Although it would have been desirable to administer a common aptitude test to all Ss, this was not feasible since course requirements prevented taking additional time from class instruction. Therefore aptitude scores were taken from the students' records. Unfortunately, the aptitude data was not complete: some Ss had only the American Council on Education Psychological Examination (ACE, 1948 edition), some Ss had only the Canadian Academic Aptitude Test (Caat), and some Ss had no aptitude test. Reliability coefficients for both the ACE and CAAT range from .80 to .90. The validity of the ACE, especially the quantitative scores, has been questioned, and since the CAAT is a relatively new test very little validity data is available (Buros, 1965, 1959).

Both the CAAT and the ACE provide a quantitative (Q), a verbal (V), and a total (T) score. However, there were no norms comparing the ACE and the CAAT, and since they differed in certain respects (the ACE developed by Thurstone and based on his theory of separate abilities, the verbal part being primarily linguistic in nature, while the CAAT was developed to meet the needs of the schools within Ontario), the scores on these two tests were treated separately. When the Ss within each experimental treatment were divided according to aptitude tests, only five divisions were large enough

to use in statistical analysis. Correlations between aptitude and the dependent variables for these divisions are given in Appendix P.

It was hypothesized that quantitative ability would be related to the structure and transfer scores for individuals within the D treatment. This relationship did occur for the O-D treatment with the CAAT test. Time was expected to correlate negatively with verbal ability and achievement was expected to correlate positively with verbal ability and general aptitude. None of these correlations occurred. Several other significant relationships existed, but it was difficult to interpret these findings because of the small sample size and the low validity of the aptitude tests themselves.

In general, few background variables correlated with transfer, leaving most of the transfer variance unexplained. Evidently the traditional variables of sex, age, and aptitude do not adequately predict individuals' ability to transfer and apply knowledge to new situations.

In this chapter the major results and minor results have been presented with discussion of the minor results. Discussion of the major results is in Chapter VII. Chapter VI presents a structural analysis of the substructure and transfer items with a discussion of the importance of such analyses for the classroom.

CHAPTER VI

ANALYSIS OF THE COMPREHENSION AND RETENTION OF STRUCTURAL RELATIONSHIPS

The analysis presented in this chapter will illustrate the type of diagnostic information that can be obtained from a test based upon diagrams as representations of structure. Such diagnostic information can not usually be obtained from the ordinary classroom test. Therefore one purpose of this chapter will be to emphasize the advantages of tests derived from a structural representation of subject matter.

The other purpose of this chapter will be to present and discuss the differences between the structural relationships which the Ss learned and remembered, and those relationships which actually existed within the reliability passage used in the present study. Each substructure was analyzed for the relationships that existed for Ss on acquisition and retention. Responses which were the same on both acquisition and retention (referred to as consistent responses) were also examined. The transfer items were analyzed similarly. The data on the six substructures is presented first, followed by the transfer data, and finally an interpretive summary of the major findings.

The following presentation of data will be used throughout the analysis of structure and transfer items. Tables presenting both acquisition-

retention and consistency data for the substructures and transfer are given. When the correct response was not dominant within the acquisition-retention responses or within the consistent responses, the percentages for the dominant one(s) are presented after the percentages for the correct response. In all cases the results are presented in percentages. The detailed analysis of each item is not given in this chapter but can be found in Appendix G. The items for each analysis are numbered according to this appendix and are coded as St#. Since the O and R groups did not differ in performance on the substructures, the results are given for the three treatments with data pooled across groups. The results for the three treatments were not pooled in order to illustrate the similarity of Ss' responses across treatments.

Structure

In order to give a perspective on the relative difficulty of the relationships within each substructure, Table 6 gives the percentages of Ss who had all the correct relationships within each substructure on acquisition and retention.

Using these percentages as a measure of difficulty, Sb6 was the easiest, then Sb1, Sbs 2 and 5, and Sbs 3 and 4. This order of difficulty was the same across treatments. Although the two easiest substructures, 1 and 6, were not exceptionally easy, i.e., not 80-90%, the remaining four substructures were extremely difficult, as indicated by the fact that in most cases no S had all the correct relationships.

Table 6
Percentage of Subjects with Perfect Substructure
on Acquisition and Retention

	NR	V	D
Sb1	28-32*	28-24	25-38
Sb2	02-00	04-02	00-00
Sb3	00-00	00-00	00-00
Sb4	00-00	00-00	00-00
Sb5	00-00	06-07	16-10
Sb6	40-44	56-61	43-45

* Second figure is retention percentage

Substructure 1: Degree of Reliability and its Practical Importance

A total of twelve relationships was tested with three true-false items (St 20, 21 and 22a). St 20 tested for six relationships, and St 21 and St 22a each tested for three relationships. Table 7 gives the acquisition-retention and consistency results for the total substructure and each of these items.

The overall difficulty level for this substructure was fairly low (20-30%). In order to be significantly greater than chance (at $p = .05$, chance = 12.5%), the percentage of correct responses should have been greater than 24%. For most treatments this was the case. The percentages

Table 7

**Substructure 1: Acquisition-Retention and
Consistency Percentages for Structures
20, 21, and 22a**

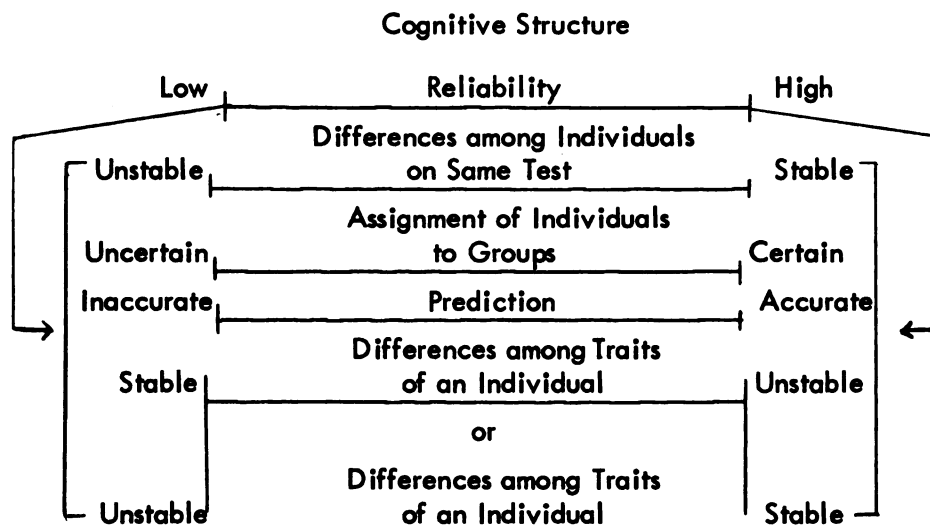
	NR	V	D
Acquisition-Retention			
Sb 1: Total	26-34*	28-23	25-38
St 20	65-76	75-58	74-85
St 21	63-53	62-57	53-60
St 22a	88-93	78-89	87-91
Consistency			
St 20	75	82	73
Correct	72	73	94
St 21	64	70	65
Correct	56	58	55
St 22a	78	83	85
Correct	100	92	95

*Second figure is retention percentage.

for each part were higher than for the total, with St 22a being the highest. In general on St 20 and 22a these percentages were significantly greater than chance (chance = 50%, at $p = .05$, greater than 68%). However, the percentages for St 21 were below this level. Differences between treatments for each item did not occur. There was no consistent trend over retention,

some figures were higher than acquisition and some were lower.

The level of most percentages on acquisition-retention indicated that this substructure was grasped by most of the Ss. Within this substructure the prediction relationship (St 22a) was grasped most clearly and the differences among traits of an individual (St 21) grasped least clearly. This trend could be explained, in part, by the nature of the items rather than by the nature of the relationship itself. The application situation in the true-false statement used for St 21 was more confusing than the situation described in St 22a. Thus the structural relationship which the Ss had reversed was differences among traits. Diagrammatically, a representation of their cognitive structures would be as follows: (The two different relationships on the last line indicate that about 50% of the Ss had the correct relationship between degree of reliability and differences among traits of an individual and the other 50% had it reversed).



The consistency level for items 20 and 22a was significantly greater than the chance level of 50% (at $p = .05$, greater than 68%). However, as with acquisition-retention, St 21 percentages were generally not significantly different from chance. Differences between treatments on each item did not occur. The percentage of correct-consistent responses was generally quite similar to the proportion of correct responses that occurred on acquisition-retention. In all cases this correct response was dominant.

Using consistency data as an indication of stability of Ss cognitive structures, only one of the four sets of relationships within Sb 1 seemed to be rather unstable (differences among traits of an individual). This relationship was also the most difficult for the Ss. Whether this difficulty and lack of consistency was the result of the item itself or the underlying relationships could only be determined by writing other items testing the same relationships.

Substructure 2: Definition of Systematic Factors and Variation and of Unsystematic Factors and Variation

A total of twelve relationships were tested by a pattern type true-false item (St 2). Scores for response patterns could range from 12 to -1 (indicating contradiction within answer pattern itself). Several scores were dominant on acquisition-retention (9, 3, and -1). These scores, and the maximum score of 12 are given separately in Table 8 but the remaining scores are grouped. Table 8 also presents the consistency data with the percentages of the two dominant consistent responses (scores of 9 and -1).

Table 8

**Structure 2: Acquisition-Retention and
Consistency Percentages for Structure 2**

	NR	V	D
Acquisition-Retention			
Score			
12	02-00*	04-02	00-00
9	37-29	28-36	47-33
3	15-17	18-28	21-23
-1	45-50	44-28	28-36
Others	03-04	06-07	13-10
Consistency			
Total	40	43	41
9	40	30	44
-1	58	43	32

*Second figure is retention percentage

Chance level on this item was .78% and percentages that were significantly greater than this level (at $p = .05$) had to be greater than 3.8%. Using this as a base, Ss who gave the correct pattern were, in general, responding at chance level. But Ss who responded with patterns that were scored either 9, 3, or -1 were definitely responding above the chance level. Acquisition and retention percentages were quite similar for each pattern. No differences between treatments occurred, i.e., if a pattern was low for one treatment it was also low for the other treatments.

Thus few Ss grasped all of the relationships (three on acquisition and one on retention had the correct pattern); yet dominant patterns did occur. The score of 9 indicated that Ss identified types of unsystematic variation and factors and systematic variation and factors correctly, but that they had variation as a cause of variation rather than factors alone as the cause of variation. This response indicated a problem in understanding causal relationships. The score of three indicated that only part of this 9 pattern held; Ss identified only one type of unsystematic variation and factors correctly. They had this type of unsystematic factor as a cause of variation but also had this type of unsystematic variation as a cause of variation. With this pattern it was impossible to infer Ss knowledge of systematic variation and factors. The -1 scores indicated a pairing of systematic and unsystematic factors and/or variation. If Ss had carefully examined the true-false items, this contradiction in terms should have been apparent. However, the contradictions might have been the result of test taking behavior. If a S was doubtful about the correct answer he might have marked several alternatives "true" in the hope of getting at least one correct (Ss did not know how the items were scored).

Diagrams of the cognitive structures of Ss corresponding to the three major response patterns are given below. The actual structure of the material is also presented.

Structure of the Material

(Cause)

		Systematic Factors	Unsystematic Factors	
			Constant	Varying
Type of Score Arrangement	Orderly Pattern	Systematic Variation	Unsystematic Variation	
	Complete Lack of Order		Constant	Varying
		X		
			X	X

Cognitive Structure

(cause)

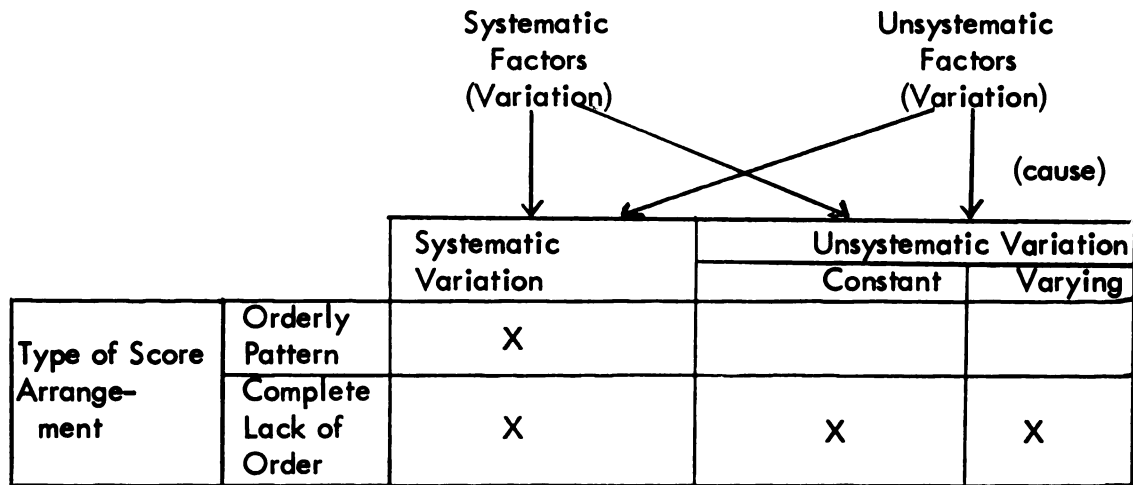
		Systematic Factors & Variation	Unsystematic Factors & Variation	
			CUF	CUV
			VUF	VUV
Type of Score Arrangement	Orderly Pattern	Systematic Variation	Unsystematic Variation	
	Complete Lack of Order		Constant	Varying
		X		
			X	X

Score of 9

		CUF	CUV (cause)
		Constant Unsystematic Variation	
Type of Score Arrangement	Complete Lack of Order	X	

Score of 3

(only constant unsystematic illustrated.)



Score of -1

(This general pattern was shown. Constant and varying unsystematic factors have been omitted from the diagram to clarify the Ss general confusion).

Considering all possible response patterns that could occur, the chance level for the consistent responses was .78%. The percentage of consistent responses that did occur was significantly higher than this level (at $p = .05$, greater than 3.8%) for each treatment. The percentages for the two dominant responses, scores of 9 and -1, were generally quite similar to the proportions that occurred on acquisition and retention. Although Ss cognitive structures did not correspond to the structure within the material itself, their cognitive structures were rather stable, as indicated by the consistency percentages. Generally speaking, those Ss who were consistent either had grasped most of the relationships or were confused.

Substructure 3: Effects of All Types of Factors

A total of 18 relationships were tested. One item, St 6, tested all of these and was a memory item. The other two items (St 4 and 5) were application items, together testing the same relationships. Analysis of the memory item will be presented first.

The memory item consisted of six descriptive parts which were to be identified as effects characteristic of systematic, constant unsystematic and/or varying unsystematic factors, or none of these factors. Thus the S could give one to three answers for each part. Table 9 gives the proportions on acquisition-retention of the correct answer, and when this response was not the dominant one the other common responses are also given. If a response was not dominant, no percentage is given for it. Each part of the item is identified by the correct answer for that part (S-Systematic, CU-Constant Unsystematic and VU-Varying Unsystematic).

Strictly speaking the chance level for each part of St 6 was 6.24% ($1/2^4$). However this was considered too conservative a chance level because Ss rarely answered with 'none of these'. Therefore a more appropriate base level of 12.5% ($1/2^3$) was considered. Yet a response bias occurred. Ss tended to respond with only one factor, i.e., the most common responses on each part were generally VU, CU, or S. Responses with more than one factor were quite infrequent. If the 12.5% level had been used for each part, significant differences would have occurred where the proportion of responses

Table 9

**Substructure 3: Acquisition-Retention Percentages
for Structure 6 - Memory Item**

		NR	V	D
Total		00-00*	00-00	00-00
Part	Response			
a S	S	67-52	48-44	57-51
	CU	14-21	17-27	17-27
	CU, S	08-15	14-04	13-09
b VU	VU	32-30	36-42	19-20
	None	29-13	16-17	20-13
	CU	20-25	18-09	08-25
	S	05-04	08-10	28-05
c CU, S	CU, S	12-12	07-07	07-08
	CU	58-28	37-31	32-39
	S	19-40	33-48	43-42
d S	S	48-31	34-29	29-34
	CU	20	18-21	25
	None	12-20	13-15	21-09
	VU	02-25	08-07	25-17
e VU	VU	65-72	72-57	51-48
	CU	20-11	16-16	16-16
	VU, CU	06-08	04-09	06
	S		05	16
f VU, CU	VU, CU	08-07	04-02	01-08
	VU	57-56	57-39	59-44
	CU	08-30	11-36	13-26

* Second figure is retention percentage.

was probably at a chance level (Ss choosing between VU, CU, and S). Therefore on the parts where only one response was required, the chance level of 33% (1/3) was used in evaluating the proportion of correct responses. On the parts where two or more responses were required the chance level of 12.5% was used in evaluating the proportion of correct responses. The same rationale was applied to determine the significance of wrong answers – 33% for one response and 12.5% for two or more responses. This rationale was applied to all other items of this type as well.

No S received the maximum score for this item as a whole on acquisition or retention. Thus Ss were responding at chance level (1.4%) and did not understand all of the structural relationships tested by the item. The data showed no differences among the treatments, but rather consistent response patterns instead. In other words, if the item was missed by the majority of Ss in one treatment it was also missed by a majority of Ss in another treatment. In addition the answers which were most common for one treatment also tended to be most common across the other treatments.

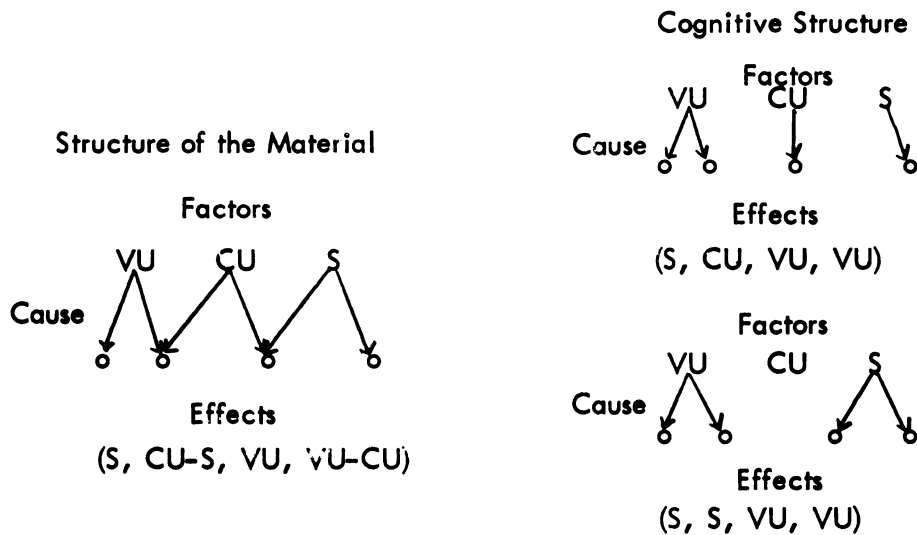
Examination of each part of the item also indicated certain trends on acquisition and retention. Parts a and d both tested knowledge of the effects of systematic factors. However part a was easier than part d (percentages for part a were significantly greater than chance, while percentages for part d were at chance level, at $p = .05$, greater than 50% required). Both parts were memory. Thus one would assume that if a S knew the correct

answer to part a he would also have answered part d correctly (d was an instance of the general case cited in part a). However, that was not the case. Parts b and e both tested knowledge of the effects of varying unsystematic factors. As with the systematic factors, there was a difference in difficulty level for these two parts, part b being more difficult than part e (percentages for part e were significantly greater than chance). Both parts were memory. In this case part b was an instance of the general case cited in part d. Thus if part d was answered correctly, part b should have also been answered correctly. This was not the case. So with both systematic and varying unsystematic factors a similar trend existed. Ss, within the memory domain, had less difficulty with the general case than with a specific example of it. The general cases (parts a and e) were taken verbatim from the structural repetitions given within the test itself.

The remaining two parts of this item (c and f) tested, in a verbatim fashion, the students' knowledge of the overlap between the effects of two factors, CU-S and VU-CU respectively. In both instances the same pattern occurred. Very few Ss understood the overlap completely; they did not grasp what was common to both factors. Instead Ss responded with only one of the correct factors. For both parts Ss were responding at chance level (11.5%) with the correct response. On part c both the CU and S responses were generally not significantly different from the chance level (at $p = .05$, greater than 50% required). On part f the proportion of VU responses was

generally significantly greater than chance level, while the proportion of CU responses was at chance level.

Considering the responses to the verbatim parts of St 6 (a, c, e, f) Ss understood the unique effects of VU and S but not the overlap in effects which each of these two factors has with CU. As with Sb 2, Ss had trouble with causal relationships. The difference between the structure of the material and the students' cognitive structures can be briefly diagrammed as follows (here only two response patterns to these four parts are illustrated - acef: S, C U, VU, VU, and S, S, VU, VU).



In conclusion Ss tended to simplify the actual structure of the material both on acquisition and retention. This tendency may be, in part, strongly related to the causal relationship itself, i.e., that individuals usually think in terms of single rather than multiple causation.

The analysis of substructures 1, 2, and 3 illustrates the type of diagnostic information which can be obtained from a structure relationships test. Such information is not obtained from the ordinary achievement test because the items do not systematically test for all structural relationships. The confusions and errors that students make on such a test are revealed by the wrong alternatives they mark. However, rarely are such alternatives specifically designed to spot structural confusions that the students might have acquired. More important, perhaps, is the fact that most tests are not analyzed for the diagnostic information that they could provide. However, a test based upon an analysis of the structure of the material is deliberately constructed such that one of the main objectives of the test is to provide diagnostic information about the learner for the tester, teacher, or researcher.

The value of this type of information is clearly illustrated by the cognitive structure of Ss on Sb 3. Sb 3 examined multiple causation relationships, a type of relationship found in many disciplines. Ss had difficulty with the relationships, and tended to simplify, not complicate, them. With this type of information a teacher could easily pinpoint learning problems. However a typical test on the same topic would have tested for only one or two of these relationships. If the student responded correctly, then the teacher would have assumed that the student "knew" the material. However, as indicated by the present structural analysis this conclusion probably would have been in error for most students. Since these relationships were prerequisites for many

of the transfer items, a complete assessment of each of them is important for predicting and interpreting performance in transfer situations.

Table 10 gives the consistency percentages for St 6. In general the degree of consistency for each part of St 6 was not above chance. Generally 40% or less of the Ss had the same answers on acquisition and retention. Within each part, the same answer was dominant for all treatments. The only major exception was part c where there was a split between CU and S as consistent answers. For parts b, c, and d the proportion of dominant consistent responses was similar to their proportions on acquisition-retention. However, for parts a, e, and f the dominant response on acquisition-retention was weighted higher as a consistent response. The abbreviation "Con" will be used on all tables to stand for the percentage of consistent responses on each part or each item.

Degree of consistency for the parts to this item was not high, indicating rather unstable cognitive structures. This degree of instability and the degree of difficulty were not expected on this memory item. In general those responses which were dominant on acquisition-retention were also the most stable responses.

On the basis of these findings one would predict that students would do poorly on application type items where the same structural relationships were the basis for the required application. The next two items described (St 4 and 5) tested this prediction. St 4 tested for effects of the three factors

Table 10

Substructure 3: Consistency Percentages for Structure 6

		NR	V	D
Part	Response			
a S	Con	42	31	41
	S	80	56	80
	CU		12	
b VU	Con	36	30	25
	VU	44	52	48
	None	27		19
c CU,S	Con	41	36	26
	CU	58	43	45
	S	34	52	48
d S	Con	34	27	33
	S	54	50	45
	VU	20	12	
	None		15	12
e VU	Con	49	34	30
	VU	100	100	80
f VU, CU	Con	34	33	42
	VU	100	80	81

on different occasions and St 5 tested for effects of these factors on the same occasion. Table 11 gives the acquisition-retention percentages for both St 4 and 5.

The acquisition-retention percentages for the correct response to all relationships within St 4 were significantly greater than chance for each

Table 11

Substructure 3: Acquisition-Retention Percentages for
Structures 4 and 5

		NR	V	D
St 4				
Total		48-36*	45-44	51-42
Part	Response			
VU	VU	84-87	84-86	80-76
	S	08-05	10-09	10-11
	CU	06-07	04-02	06-08
CU	CU	62-47	60-62	59-63
	VU	20-12	18-22	21-20
	S	16-36	21-18	13-10
S	S	64-55	76-73	73-74
	CU	22-28	18-18	27-25
	VU	05-04	04-07	
VU	VU	80-83	82-84	90-91
	CU	12-09	18-17	08-04
	S	04		02-02
St 5				
Total		02-02	00-00	00-00
Part	Response			
CU,S	CU,S	04-02	00-00	00-04
	CU	52-36	32-41	40-39
	S	33-55	62-48	50-46
	VU	08-04	02-06	06-04
VU	VU	88-75	78-70	73-73
	CU	09-26	18-24	17-19
	S		04-02	05-04
CU	VU	40-25	44-22	24-21
	CU	24-37	37-32	38-40
	S	38-32	21-44	27-28

* Second figure is retention percentage.

treatment ($p = .05$, greater than 4.9%, chance level = 1.2%). However the acquisition-retention percentages for the correct response on all parts of St 5 were not significantly different from chance for each treatment (chance = 1.4%). The high percentage of correct responses on St 4 was not expected, because Ss had found St 6, the memory item which tested for the same relationships, very difficult. However, the difficulty of St 5 was consistent with the results on St. 6. As with St 6 the most common responses were either VU, CU, or S.

The percentages for the correct responses for each part of St 4 were significantly greater than chance for each treatment (at $p = .05$, greater than 50%, chance = 33%). Each part required only one response. This requirement alone may have increased the percentage of correct responses. However, the level of percentages for similar parts of St 6 were not as high. Perhaps the technical wording in the memory item tended to confuse Ss in classification, whereas in the application item the labels themselves and their apparent meaning served as strong cues to the appropriate classification.

For St 5 the first part was quite difficult (responses were at chance level of 12.5%). The correct answer for this part required both CU and S. In general Ss tended to respond with only one of these answers. This response pattern was consistent with results on St 6 (part c). The other two parts both tested for application of VU factors. However, the first part was easier than the second (first was significantly greater than chance but the second was

generally not). An explanation for this is not given, but it does indicate that two application situations are not of equal difficulty for Ss even though the same principle is required in both. The double response requirement on St 5 and the response bias of the Ss could, in part, explain the low percentages that occurred. However, the complexity of the structure itself was probably another factor.

On the basis of these items it was difficult to determine exactly what structural relationships the Ss acquired. But the unique part of the VU and S effects were acquired by most of the Ss. The CU effects overlapped with VU and S effects, and Ss acquired only half of this overlap in each case.

Consistency data for the total patterns on both St 4 and St 5 was significantly greater than chance levels of 2.9% and 6.9% respectively (Table 12). Since the correct pattern was the most consistent one on St 4, separate figures are given for it. Two patterns were high on consistency for St 5 and the percentage figures are given for both of these patterns. Consistency data for the parts of each item is also presented.

Consistency was generally higher for each part on St 4 and 5 than for the total pattern. This finding was similar to the consistency data on the memory item, St 6. Also the degree of consistency, in general, tended to be higher on these items than on the memory item. At least 50-60% of the Ss were consistent on St 4 and 5, while the memory item had at least 40%. The exception to this trend was the third part of St 5 where the consistency was

Table 12

Substructure 3: Consistency Percentages for
Structures 4 and 5

		NR	V	D
St 4				
Total				
	Con	24	42	31
	VU, C U, S, VU	72	62	54
Parts				
	VU	78	75	72
	CU	51	62	49
	S	62	64	64
	VU	70	84	82
St 5				
Total				
	Con	17	21	21
	CU, VU, S	50	36	33
	S, VU, CU	13	42	32
Parts				
	CU, S	48	66	49
	VU	72	69	65
	VU	32	36	39

lower. If difficulty is related to consistency, i.e., difficult items yielding fairly low consistent results, the relatively high consistent percentages on St 4, in contrast to St 6, can be explained. However St 5 was as difficult as St 6 but St 5 responses were more consistent. Another factor may have been the technical wording used in St 6. Again no differences between treatments occurred. If a pattern was low (or high) for one treatment it was also low (or high) for the others.

The consistency data was in accord with the basic acquisition-retention data on both the memory and application items, with the application items (especially St 4) being easier and more consistent. Both of these trends were contrary to expectations. Attempted explanations of these results have been given.

Substructure 4: Definition of Reliability, Reliability Coefficient, and Correlation Coefficient

A total of 21 relationships were tested by using six different items. Two items (St 14, and 15) tested for 16 of these relationships. St 13 was a pattern true-false item (two statements); St 14 was a pattern true-false item (six statements); St 15 was two independent true-false items; St 16 and 18 were true-false items; and St 19 was a matching item (one stem and two correct options). Acquisition-retention and consistency percentages are given in Table 13. If the correct response was not the majority response, the dominant responses are also given.

Two items (St 13 and 14) had very few Ss answering correctly. For St 13 Ss were responding significantly below the chance level of 25% (at $p = .05$, less than 10.6%), while for St 14 Ss were responding at the chance level of 1.6%. The error in St 13 was that Ss interpreted the correlation coefficient as a cause, as causing high or low relationships between tests rather than describing such relationships. The error in St 14 was that Ss stated that reliability was a quantitative index. The substructure itself made a

distinction between the concept of reliability and the quantitative index of the reliability coefficient.

St 16 generally exhibited a 50-50 split between the two possible responses (a true-false item). This was a rather straight forward interpretation of the relationship between parallel and non-parallel tests and reliability. However, Ss responded at chance level. Responses to St 18 (true-false item) indicated that Ss did not integrate the concepts of parallel tests and unsystematic variation with the degree of reliability. In fact, responses to St 18 were significantly below the chance level of 50% (at $p = .05$, less than 31%).

Table 13

Substructure 4: Acquisition-Retention and
Consistency Percentages for All Items

		NR	V	D
Acquisition-Retention				
Total		00-00	00-00	00-00
Item	Score			
St 13	3	00-00	00-00	00-00
	2	76-85	76-73	83-86
	0	23-13	18-24	17-14
St 14	10	00-02	02-02	03-02
	9	80-74	74-75	76-79
St 15a	2	66-60	69-74	63-64
15b	4	67-70	57-64	83-57
St 16	2	55-79	57-41	42-52
	0	45-21	43-49	58-48
St 18	1	20-15	29-12	17-09
	0	80-85	71-88	83-91
St 19fd	2	58-66	88-74	71-59

Table 13 (Continued)

		NR	V	D
Item	Score			
Consistency				
St 13	Con	66	75	87
	2	93	83	94
St 14	Con	70	70	60
	9	97	94	100
St 15a	Con	61	79	71
	2	75	71	69
15b	Con	66	52	59
	4	81	70	86
St 16	Con	64	59	61
	2	77	62	12
	0			57
St 18	Con	80	72	81
	0	89	87	97
St 19	Con	59	70	55
	2	76	97	84

*Second Figure is retention percentage.

St 15 a and b checked if Ss knew the relationship of unsystematic variation to reliability and correlation coefficients, and if they knew the relationship of parallel and non-parallel tests to both of these coefficients. Generally speaking Ss were responding at chance level (50%) on both of these parts (at $p = .05$, greater than 68% required). However some treatments

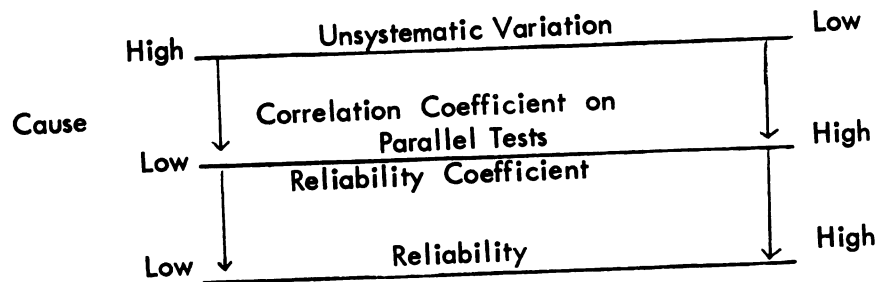
were above chance level - for a, V and for b, NR on retention and D on acquisition. St 19 tested for two aspects of reliability, its relationship to unsystematic factors and its strength. Using the chance level of 25% (probability that Ss would respond with both of these factors correctly, ignoring the other options in the matching item which were relevant to other stems), Ss were responding significantly above chance (at $p = .05$, greater than 39%).

In general, most of the relationships (10-16 out of 21) within substructure 4 were acquired and retained at a fairly high level. Again no differences among treatments existed. Diagrammatic representation on the subject matter and the Ss cognitive structures is presented below. The question marks (?) indicate that differences between treatments existed. The major errors were related to the causal relationship and precision of definitions. The latter error would imply that Ss find it difficult to distinguish clearly among closely related concepts.

The degree of consistency was fairly high for each item. Degree of consistency was significantly greater than chance for Sts 13, 14, 18, and 19. For St 15a two of the three treatment percentages were significantly above chance level. For St 16 and 15c the responses were generally at chance level. In all cases the response that was most common on acquisition and retention was also the most consistent. Generalizing across all items on substructure 4, the cognitive structures of the Ss were rather stable.

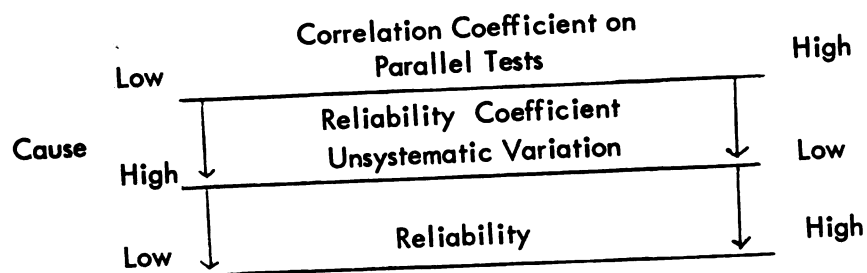
Structure of the
Material

		Rel.	Rel. Coeff.	Corr. Coeff.
Quant. Index			X	X
Degree of Unsys. Var.		X	X	X
Tests	Par.	X	X	X
	Not Par.			X



Cognitive
Structure

		Rel.	Rel. Coeff.	Corr. Coeff.
Quant. Index		X	X	X
Degree of Unsys. Var.		X	X	X
Tests	Para.	X	?	?
	Not Par.	X	?	?



Substructure 5: Parallel forms versus Parallel Tests

A total of six relationships were tested by a matching item (St 19). The matching item was a memory one consisting of two parts, the characteristics of parallel tests and the characteristics of parallel forms of tests. The percentage of Ss who responded with the correct pattern and dominant incorrect patterns for both parts on acquisition and retention is given in Table 14. In addition the percentage of correct patterns for each response to the parallel forms and parallel tests parts is given.

In contrast to the other substructures, definite differences between treatments appeared to exist on the total substructure, with the D treatment having the highest percentage of correct patterns followed by V and NR. Chance level for the total substructure was 1.5% with 5.75% required to be significantly higher than chance at the .05 level. Therefore the NR treatment was always below chance while the D treatment was always above chance, with the V treatment at chance level. However a Chi-Square test showed that the only significant differences between treatments occurred on acquisition (Chi-square = 11.4, significant at $p < .01$).

The order of treatment percentages was also repeated with the parallel tests part and to some extent with parallel forms. For the parallel forms part most percentages were significantly above the chance level of 12.5% (at $p = .05$, greater than 23%). However, this was not the case for the parallel tests part. On retention all percentages were at the chance level,

Table 14

**Substructure 5: Acquisition-Retention Percentages
for Structure 19 - Parallel Forms and Tests**

	NR	V	D
Total	00-00*	06-06	18-10
Parallel Tests			
Response Pattern			
ab**	04-04	18-14	32-14
c	14-19	17-16	12-13
bc	18-16	12-09	09-17
abc	29-25	19-29	24-26
b	05-08	21-12	06-02
Correct Response - Each Characteristic			
a	43-42	47-58	67-52
b	69-61	73-62	73-68
c	25-31	65-24	49-29
Parallel Forms			
Response Pattern			
bc**	27-22	30-33	43-29
b	22-23	12-09	14-18
c	15-17	19-18	14-28
abc	07-09	16-16	
Correct Response - Each Characteristic			
a	71-66	66-67	70-62
b	66-66	62-62	74-60
c	68-63	70-72	70-61

* Second figure is retention percentage

** Correct response pattern.

while on acquisition some treatments were above and some at the chance level.

A Chi-square test for the acquisition percentages on the parallel tests part

showed that the treatments differed (Chi-square = 29.3, significant at $p < .001$).

The differences found between the treatments were consistent with the basic theoretical predictions behind the study. However, since the pattern was not replicated on any of the other substructures, the differences were probably a chance occurrence. No wrong pattern was predominant within or across treatments on the parallel tests or forms parts, even though the correct pattern was not given by a great proportion of Ss.

Within the parallel tests and parallel forms parts, the percentages of correct answers for each option was fairly high. Using 68% as the cutting point for significance (at $p = .05$) for the chance level of 50%, option b tended to be significantly higher than chance level for parallel tests, while the other options tended to be at the chance level. For the parallel forms part, options a and c tended to be significantly above chance. Responses on option b were generally at the chance level for both groups.

On the basis of these results it was difficult to generalize about the cognitive structures of the Ss. However, using the percentages for the NR treatment on each option, the differences between the structure of the material and the cognitive structure of the students can be diagrammed as follows:

Structure of the Material

Meet Stat. Crit.	Meas. Same Trait	Always Sim. in Content	
	X	X	Parallel Forms
X	X		Parallel Tests

Cognitive Structure

Meet Stat. Crit.	Meas. Same Trait	Always Sim. in Content	
	X		Parallel Forms
	X	X	Parallel Tests

Table 15 gives the consistency percentages for this substructure.

Table 15

Substructure 5: Consistency Percentages for
Structure 19 - Parallel Forms and Tests

	NR	V	D
Total	19	16	16
Correct	00	13	62
Parallel			
Tests	39	36	31
Correct	03	12	39
a	72	70	61
b	68	73	64
c	78	65	66
Parallel			
Forms	27	48	31
Correct	42	35	50
a	68	68	74
b	59	71	58
c	61	73	56

There were few consistent patterns for the total structure, although all percentages were above the chance level of 1.5%. Few of these consistent patterns were the correct ones. The degree of consistency was higher (significantly different from chance level of 12.5%) for the two parts of parallel tests and parallel forms. Even higher percentages of consistent responses for each option occurred with about half of these percentages above chance. The proportion of consistent-correct responses for parallel tests and parallel forms tended to vary with each treatment. Thus even though the absolute level of the consistency percentages was rather low, most of them were above chance, indicating that Ss cognitive structures were rather stable for the relationships within this substructure.

Substructure 6: Methods of Estimating Reliability Coefficients

Twelve relationships were tested by a matching item (St 7-1st) which consisted of five parts. One set of relationships was tested twice (parts a and e of the item). The percentage of correct patterns on acquisition-retention and the percentage of consistent responses are given in Table 16.

All acquisition-retention percentages were significantly greater than the chance level of 3.1% (at $p = .05$, greater than 8.8%). In fact this substructure appeared to be acquired and retained the best of any of the substructures. As was true with the other substructures, the parts showed a higher proportion (on an absolute level) of correct responses than the total structure.

Most of these percentages were significantly greater than chance (50%, at $p = .05$, greater than 68%). In general there did not seem to be much difference in difficulty between parts a and e. No consistent differences between treatments occurred. For this substructure the diagrams of the students' cognitive structures would be the same as the structure of the material.

Table 16
Structure 6: Acquisition-Retention and Consistency
Percentages for Structure 7- 1st

	NR	V	D
Acquisition-Retention			
Total	40-41*	55-61	43-46
Parts			
a	60-75	75-80	69-73
b	69-77	97-88	86-79
c	76-89	90-93	88-89
d	78-78	80-88	74-82
e	72-76	74-86	90-76
Consistency			
Total	39	46	54
Correct	94	96	65
Parts			
a	78	86	82
Correct	78	78	61
b	80	91	82
Correct	79	92	71
c	84	90	85
Correct	90	98	67
d	75	81	81
Correct	91	95	71
e	72	72	77
Correct	82	94	75

*Second figure is retention percentage.

In general the percentage of consistent patterns for the total structure was lower than the degree of consistency for each part. However, both the total structure and part percentages were significantly greater than chance (3.1% and 50% respectively). The percentage of consistent-correct total and consistent-correct part patterns was high across most treatments. These high consistency figures indicated that Ss cognitive structures for this substructure were quite stable.

Transfer

Seven of the ten transfer items were based upon a relationship between Sbs 3 and 6. The remaining three transfer items were based on other substructures. The data for these three transfer items is presented first.

Transfer based on Substructure 1

This true-false transfer item (St 22b) was based upon the generalization of the importance of reliability to a correlational situation. The percentage of correct and consistent responses is given in Table 17.

This transfer item was fairly easy. Most acquisition-retention percentages were significantly greater than chance (chance = 50%, at $p = .05$, greater than 68%). The percentage of consistent responses was also fairly high with the majority of these responses being the correct one. The relatively high level of performance by Ss on this transfer item was in accord with the level of performance by Ss on Sb 1.

Table 17

Transfer, Substructure 1: Acquisition-Retention and
Consistency Percentages for Structure 22b

	NR	V	D
Acquisition- Retention	70-55	77-75	80-80
Consistency			
Total	68	68	82
Correct	69	89	86

* Second figure is retention percentage.

Transfer based on Substructure 2

This transfer item (St 3) was based upon the effects of the simultaneous occurrence of unsystematic and systematic factors. It was a multiple true-false item yielding a pattern of responses. Table 18 gives the percentage of correct and dominant responses for acquisition-retention and consistency.

The proportion of correct responses was significantly greater than chance (chance = 6.25%, at $p = .05$, greater than 14%). However the dominant response of -1 was also significantly greater than chance for all treatments, and the absolute level of this percentage was greater than that for the correct response. Thus Ss seemed to contradict themselves, marking two alternatives "true" which were logically impossible to exist under the conditions of the question. The degree of consistency was also significantly greater than chance (6.3%), with the proportion of correct and contradictory

patterns being similar to their corresponding proportions within the acquisition-retention data.

Table 18
Transfer, Substructure 2: Acquisition-Retention
and Consistency Percentages for Structure 3

	NR	V	D
Acquisition-Retention			
Correct	16-23*	18-20	20-28
-1	70-62	67-62	60-56
Consistency			
Total	62	54	55
Correct	14	19	24
-1	83	81	65

* Second figure is retention percentage.

This item was more difficult than originally expected, even though Ss were generally responding correctly above the chance level. The heavy proportion of contradictory response patterns indicated that Ss were confused. Very few Ss responded correctly to all relationships within substructure 2 itself, and contradictory patterns occurred on this substructure. Thus poor performance on this transfer item was consistent with findings on Sb 2.

Transfer based on Substructure 5

This true-false transfer item (St 17) was based on an implication made from the fact that statistical criteria is a characteristic of parallel tests. Table 19 gives the acquisition-retention and consistency percentages for this item.

Table 19

Transfer, Substructure 5: Acquisition-Retention
and Consistency Percentages for Structure 17

	NR	V	D
Acquisition-Retention	64-61	59-64	73-67
Consistency			
Total	55	72	55
Correct	73	65	72

*Second figure is retention percentage.

The proportion of correct responses was generally at chance level (50%) across treatments on acquisition and retention. The consistency percentages were generally also at chance (chance = 50%, at $p = .05$, greater than 68%). The proportion of consistent-correct responses was similar to the corresponding proportions within the acquisition and retention data. The percentage of correct responses on the item testing for the basic structural

relationships used in this transfer item (Sb 5, St 19 parallel tests, part a) was also at chance level. Again the results supported the hypothesis that Ss would not perform highly on transfer items unless they had performed highly on the structural relationships which were the basis for the transfer.

Transfer based on Substructures 3 and 6

Six transfer items (St 7-2nd, 8, 9, 10, 11, 12) were based on the same structural relationships in Sb 3 and 6. The basic relationships tested, in one form or another, by these transfer items dealt with the transfer structure diagrammed in Appendix G.

Structure 7-2nd.--Item 7-2nd directly tested for each of the relationships within the transfer structure (parts a and e were duplications). Table 20 gives the acquisition and retention data for this item on both the total structure and each part.

Using the total response pattern as the criterion, this item was quite difficult (Ss responded at the chance level of .058%). Only one S answered with the correct pattern. Using 33% and 12.5% as the chance levels for parts ade and bc respectively, Ss responded at the chance level for both parts. The most common responses made by the Ss to all five parts were rather evenly divided among VU, CU, and S.

The difficulty of the item and its separate parts may have been, in part, a result of response bias. The common tendency by Ss was to indicate

Table 20

Transfer, Substructures 3 and 6: Acquisition-Retention
Percentages for Structure 7-2nd

		NR	V	D
Total		00-00*	00-02	00-00
Parts	Response			
a VU	VU	24-21	20-25	13-17
	S	43-49	39-53	51-43
	CU	19-20	31-14	26-27
b VU, CU, S	VU, CU, S	04-04	00-07	02-04
	CU	34-47	35-20	18-47
	VU	28-25	35-38	44-27
	S	19-13	12-17	24-18
c VU, CU, S	VU, CU, S	05-04	04-06	00-02
	VU	19-08	21-15	25-17
	CU	23-32	33-41	32-36
	S	40-44	34-29	29-36
d VU	VU	28-35	23-25	25-24
	CU	38-17	23-30	35-40
	S	33-32	45-29	33-28
e VU	VU	28-34	35-32	13-23
	CU	36-13	31-22	27-30
	S	35-40	23-27	43-41

* Second figure is retention percentage

only one factor. The fact that parts b and c required all three factors as the correct response reduced considerably the absolute proportion of Ss who answered the item correctly. This transfer item was an integration of Sbs 3 and 6. Sb 3 was not grasped by any of the Ss, however Sb 6 was understood by the majority. The difficulty of this transfer item then could also reflect the difficulty Ss had with Sb 3. On the basis of these results it was difficult to generalize about the cognitive structures of the Ss. However, using the percentages for the NR treatment, the cognitive structure of the students can be diagrammed as follows (structure of the material is indicated with circles "O"):

Cognitive Structure

Systematic Factors		X O			O
Unsystematic Factors	Const.	O	X		X O
	Vary.	O	O	X O	O
		Test Retest	Internal Consistency	Parallel Forms Immediate	Parallel Forms Delayed

Consistency data for the total transfer structure and its parts is given in Table 21. The degree of consistency for the total pattern was generally low in absolute terms. However for the NR and V treatments, the degree of consistency was significantly greater than chance (.058%). The degree of consistency for each part was higher on an absolute level than for the total structure, but no consistent results regarding significance occurred. For some

Table 21

Transfer, Substructures 3 and 6: Consistency
Percentages for Structure 7-2nd

		NR	V	D
Total		10	04	00
Parts	Response			
a VU	Con	52	36	54
	VU	26	31	15
	CU	18	17	26
	S	54	45	59
b VU, CU, S	Con	21	36	33
	VU, CU, S	19	00	00
	VU	17	50	38
	CU	40	19	34
	S	21	19	27
c VU, CU, S	Con	25	26	34
	VU, CU, S	15	11	00
	VU			16
	CU	37	23	22
	S	47	23	62
d VU	Con	27	48	27
	VU	15	32	18
	CU	36	20	56
	S	45	43	25
e VU	Con	37	50	23
	VU	48	41	19
	CU	17	23	22
	S	31	31	60

treatments the percentages were significantly above chance and for others they were not (chance for ade = 33%, chance for bc = 12.5%, greater than 50% and 23% required respectively at $p = .05$). The consistent answers for each part seemed, in general, to be divided among VU, CU, and S, parallel to the division that occurred with the acquisition-retention data.

The general trend of responses found on this transfer item was consistent with previous results. Ss performance was higher within parts of the item than on the total structure itself. Performance was generally at chance level and stability of cognitive structure was also low. The general low performance on this item was in accord with the high difficulty of Sub-structure 3.

Structure 8.--As stated before St 7-2nd tested the Ss knowledge of the types of score variation that could be distinguished in the four situations used for estimating reliability coefficients. St 8 used this same basic transfer structure, but also required Ss to integrate the definition of the reliability coefficient. This transfer item asked Ss to list the types of factors which affect the reliability coefficients obtained from the four different estimation methods. Since the reliability coefficient had been defined in terms of unsystematic variation, systematic factors were eliminated. Table 22 gives the acquisition-retention data for this item.

Only one S responded correctly on all parts of the item (chance level of .17%). In general, each part was also difficult. Ss responded at the

chance level of 12.5% for parts a and d and at the chance level of 33% for parts b and c. For each part Ss responses were divided among VU, CU, and S, and the most common response was not the correct response. Again the acquisition-retention results for this item, the differences between the absolute difficulty levels of the parts and the whole structure, and the common wrong responses were the same patterns that occurred on most other items of this type. The difficulty of the total item was consistent with the difficulty of Sb 3.

Table 22
Transfer, Substructures 3 and 6: Acquisition-Retention
Percentages for Structure 8

		NR	V	D
Total		00-02*	00-00	00-00
Parts	Response			
a VU, CU	VU, CU	09-12	00-02	04-05
	VU	37-10	53-40	37-33
	CU	09-14	16-11	15-15
	S	21-47	22-27	35-36
b VU	VU	25-34	18-25	22-23
	CU	14-21	16-14	24-22
	S	36-23	60-50	39-40
c VU	VU	08-13	14-09	08-04
	CU	47-46	39-41	48-49
	S	23-13	39-26	26-31
d VU, CU	VU, CU	09-08	00-02	00-02
	VU	35-45	52-51	54-55
	CU	16-16	26-24	17-19
	S	18-11	09-04	18-12

* Second figure is retention percentage.

Consistency data for St 8 is given in Table 23. The absolute level of the total consistency percentages was rather low, but the percentages were significantly greater than chance (.17%). The degree of consistency for parts a and d were significantly greater than chance (12.5%), but the percentages for parts b and c were at chance level (33%). The most consistent response within each part was usually not the correct response. This could reflect the fact that the correct response did not occur very frequently for the parts on acquisition and retention. The most consistent response for each of these parts was also the most common response on acquisition-retention. As mentioned before responses to all parts were divided among VU, CU, and S on acquisition and retention. However, only two or one of these responses were highly consistent. Generally speaking, the Ss were not using a very stable cognitive frame of reference.

Structures 9, 10, 11, and 12.--The remaining four transfer items each dealt with the same transfer structure and definition of reliability as did St 8. The format of these items differed from the other two transfer items. St 9 was multiple true-false item; St 10 was a true-false item; St 11 was a multiple true-false item; and St 12 was three separate true-false items. Acquisition and retention data for all these items is given in Table 24.

Table 23

Transfer, Substructures 3 and 6: Consistency
Percentages for Structure 8

		NR	V	D
Total*		12	14	06
Parts	Response			
a VU, CU	Con	36	43	46
	VU, CU	06	00	05
	VU	23	37	50
	S	42	41	25
b VU	Con	33	64	47
	VU	50	17	11
	S	26	71	62
c VU	Con	38	41	32
	VU	06	00	06
	CU	73	77	51
	S			12
d VU, CU	Con	49	46	37
	VU, CU	04	00	00
	VU	57	73	80

*Three of these responses were S, VU, CU, VU.
Four of these responses were VU, S, CU, VU.
The remaining eight responses did not overlap.

Table 24

Transfer, Substructures 3 and 6: Acquisition-Retention
Percentages for Structures 9, 10, 11, and 12

		NR	V	D
St 9		17-15*	20-14	23-18
	0	51-65	48-57	42-57
	-1	31-20	32-29	35-25
St 10		60-46	69-66	43-52
St 11		13-15	13-24	15-17
	1	33-32	26-29	39-29
	0	17-27	29-20	15-25
	-1	36-25	32-25	32-35
St 12	total	47-55	48-43	38-35
	a	60-65	71-60	50-46
	b	78-75	72-73	67-62
	c	79-92	81-84	85-90

* Second figure is retention percentage.

Although based upon the same structural relationships the transfer items varied in difficulty, with St 12 (total) being the easiest, when the chance level was considered. St 12 was the only item which was significantly greater than chance (12.5%, at $p = .05$, greater than 24%). This high level of performance might have occurred because the item parts could be answered "true" or "false" on a common-sense basis. However, if Ss had been asked for the reason for their answer, and the reason then scored in terms of Ss comprehension of unsystematic variation, the item might have been more difficult.

(percentages for each part of St 12 were significantly greater than chance).

So the apparent easiness of this item did not necessarily contradict the fact that Substructure 3 was difficult or that the other transfer items were difficult.

Consistency data for these four items is given in Table 25.

Table 25

Transfer, Substructures 3 and 6: Consistency
Percentages for Structures 9, 19, 11, and 12

		NR	V	D
St 9		48	41	49
	Correct	13	18	24
	O	61	60	55
St 10		66	62	67
	Correct	55	77	50
St 11		38	26	29
	Correct	17	7	13
St 12	Total	51	49	57
	Correct	71	51	47
	a	43	41	29
	b	43	38	41
	c	49	40	51

The degree of consistency was significantly greater than chance for St 9, St 11, and St 12. However, the correct response was not predominantly consistent for St 9 and St 11. These two items were also difficult as indicated by the acquisition-retention data. Except for St 12, the results on these transfer items were generally in accord with the difficulty and lack of consistency found on Substructure 3 and the other two transfer items

(St 7-2nd and St 8). An explanation of the easiness and consistency of St12 has been given.

Summary and Interpretation

In general, differences between treatments did not occur. In fact, when the correct response was not the dominant one, the same wrong response(s) was dominant across the treatments. Of course, this rather striking similarity among treatments did not provide support for the major hypothesis of the study, i.e., that the different versions of the reliability passage would produce different performances on the structure and transfer tests. Another rather consistent result was that the percentages of correct and wrong responses were quite similar on acquisition and retention for all treatments. Thus the Ss cognitive structures did not simplify or breakdown with time.

For most items and substructures the absolute level of correct responses was higher for the parts of the item or substructure than for the total item or substructure. However, when these percentages were 'corrected for chance', this order usually did not occur. If Ss responded at a chance level for the total item they also did so on its parts. There was a tendency for difficulty to be related to consistency, i.e., the more difficult the item, the less consistent the responses from acquisition to retention. In other words, if a student did not grasp the structural relationships then his cognitive structure was apt to be more unstable than if he did understand all the relationships.

Transfer levels of difficulty and consistency were in accord with the related substructure levels of difficulty and consistency. This result did provide support for another major hypothesis, that high performance on transfer depended upon Ss understanding the structural relationships used to generate the transfer items.

Ss tended to respond to classification type items (VU, C U, and S) with only one factor, even though two or three factors were required for the correct response. This tendency may have represented only a response bias, but it could also have represented a general cognitive tendency by Ss in the process of learning (as opposed to consolidation and review) to simplify cognitive structures. This simplification is in accord with an information-theory learning position, i.e., the effects of Ss coding systems and channel capacity, and the consequent limits on Ss ability to process new information.

Ausubel's theory of meaningful verbal learning would not directly predict these results, but simplification could be considered consistent with his viewpoint. Assuming that much of the technical material on reliability was new to the Ss, Ss would then have few relevant existing subsuming concepts for this material. However, in such situations Ausubel postulated that individuals use any concepts that might be appropriate, even though they are often inadequate. Assuming then that Ss used inadequate subsuming concepts, the very inadequacy of these concepts would probably lead to simplification of structural relationships. Inadequate subsuming concepts would then provide

relevant structures for only part, not all, of the new material. On the other hand, if the Ss did not attempt to learn meaningfully, i.e., not provide any subsuming concepts however inadequate, then Ausubel would postulate that Ss would learn rotely. In order for all relationships to be acquired in this way repetitions of the material over time are required. This type of repetition was not allowed in the present study, so Ss could only master part, not all of the relationships.

This interpretation of the simplification that occurred is also consistent with the "blueprint" theory of diagram presentation. As stated in Chapter III it was assumed that when a S encountered a diagram a relevant mediation process would be "triggered." However, if the written material was acquired in a less than adequate meaningful way or was acquired rotely, the Ss then only understood part of the content. Thus the diagram would still evoke a mediation sequence, but the sequence would be inadequate or inappropriate.

Certain types of relationships were difficult for the Ss, primarily definitional and causal. As shown in Sb 4 Ss lacked precision in making definitions, e.g., reliability as a quantitative index, parallel and non parallel tests as related to reliability, reliability coefficients, and correlation coefficients. Ss also had great difficulty with causal relationships: classifying descriptive factors as causes (as with systematic and unsystematic variation causing variation in Sb 2 and correlation coefficient causing unsystematic

variation in Sb 4) and an inability to grasp multiple-causation (VU, CU, and S each had two effects, some overlapping, in Sb 3). Ss also tended to contradict themselves (-1 scores). Although such contradictions could be attributed to careless reading and test taking behavior, the cause of such responses could be more complicated. Perhaps some Ss are unable to detect contradictions in definitions of concepts, in causal relationships, in descriptions of data, etc. The implications of these findings are discussed in the last chapter.

Several major implications for this type of analysis can be given. First, improvements in present test construction by teachers can be made using such procedures. Diagrams representing structure can be applied to all subject matter, if structure is defined as presented in this study. Such representations quickly pinpoint sequential dependencies and transfer material. Instead of the usual haphazard procedure for generating transfer items, an algorithm is provided by the diagramming procedure. The importance of diagnostic information for the teacher and researcher has already been emphasized.

Second, too often tests are used only to rank students within a class. This procedure is often a result of the tests themselves; they are not deliberately structured to yield diagnostic information. With the present emphasis in education upon individualized instruction, the need for ranking students is eliminated. Instead information on what the student has and has not learned is required. The type of test construction described in the present study provides one approach to this need.

Third, structure tests could provide information on the way in which individuals learn and retain structures. Several areas of investigation would be the number of trials necessary for all structural relationships to be acquired, the time at which transfer is readily made, and parametric data on the retention of structural relationships after acquisition, overlearning, and application. With the type of information provided by structure tests, rather extensive data on learning and retention processes could be obtained.

CHAPTER VII

DISCUSSION OF MAJOR RESULTS AND CONCLUSIONS

The major emphasis in this chapter is an explanation of the unexpected results that occurred for practically every major hypothesis. Then a brief presentation of the implications of the study is given.

Time

The major difference between Groups O and R was on the time required to read the reliability passage. The times for the experimental treatments in Group R varied less than the times for the treatments in Group O, with the D treatment requiring less time in the R group and the NR treatment in the R group requiring more time. Differences in administrative procedures could account for these results. For Group O each treatment was administered in a separate room, while for Group R all treatments were given in the same room. Perhaps group pressures for Ss in Group R, similar to pressures in the Asch (1956) line judgement studies, could have made Ss in the NR treatment hesitant to turn in their papers early and could have influenced the Ss in the D treatment to read faster in order to turn in their papers with the majority of the Ss. Such pressures were not as great for individuals in Group O because

all Ss in one room had the same version of the reliability passage, allowing a greater spread among the mean scores for each treatment than for Group R. However, similar pressures could have existed for Ss in Group O. This could partially account for the time differences among the experimental treatments within Group O itself; the group pressure yielding a smaller variance on the time scores than might be expected if the Ss were reading on an individual basis.

Achievement, Structure, and Transfer

As expected, no differences existed in achievement among the experimental treatments, and all treatments scored higher than a control group which did not read the reliability passage. However, in light of the other results, the achievement data did not offer definitive evidence to support the original hypothesis.

One reason for this conclusion is that the differences between the means for the experimental and control groups, although significant, were not as large as one would have expected if the Ss had actually comprehended the material. According to test theory, most test items should be near the 50% difficulty level (adjusted for chance). Thus the mean score on a test should be approximately halfway between the chance and the maximum score. For the achievement test this score was 19 (halfway between 7.5 and 30). However, the obtained mean score of 15 for each treatment was lower than this,

indicating that the test was too difficult for the Ss. In addition, the theoretical formulation would be supported only if all the hypotheses regarding the differences between the experimental treatments on achievement, structure, and transfer were supported. The structure and transfer hypotheses were not supported.

The structure and transfer data also implied that the Ss did not comprehend the material as well as had been anticipated. On the average the experimental treatment's scores on structure were higher than the control group's scores, although this difference was not as large as had been expected. However, no differences existed between the experimental treatments and the control group on transfer.

Thus it would appear that inadequate comprehension could account, in part, for the unexpected results. Additional support for this interpretation was indicated by the questionnaire results. For both the V and D treatments Ss indicated little awareness of the relationships between test items and review sections, verbal or diagrammatic. If the structure of the material had been comprehended (it was in effect repeated three times), more Ss within the V and D treatments would have indicated awareness of a connection between the test and the reviews. Subjects' confusion about the diagrams, as stated on the questionnaire, also indicated lack of comprehension. Without a majority of the Ss attaining comprehension, a test of the theoretical formulation was really not obtained. Apparently one reading was not sufficient for the type of performance required by the test items.

Another consistent result was no retention drop on achievement, structure, or transfer. Again this was contrary to expectation. One reason for this could be that the unique format of the test increased the memory for some items for some Ss. Secondly, perhaps one week was an insufficient period of time for memory changes to occur with meaningful material. Third, if Ss were responding near chance level a retention drop would not be expected. These reasons probably do not adequately explain the stability of the scores, but at present no other explanations are available.

Although these results were unexpected, the pilot study results were as hypothesized. On the final revisions of the materials the D Ss were performing at a higher level than the V Ss on achievement, structure, and transfer. The average scores of the V and D pilot Ss on each of these three variables were higher than the highest average scores on these variables from all the six treatments in the present study. Perhaps these differences could be partially explained by the differences in time spent reading the reliability passage. The pilot Ss who received the final versions of the materials took an average of 52 minutes to read the passage compared to the high average of 47 minutes for the six treatments in the present study. In fact, the average time for all Ss in the pilot study was slightly higher than the highest average time for the six treatments. Even though time was not correlated with performance in the present study, perhaps if Ss had read the material more slowly differences among treatments would have occurred and time might have been correlated with performance.

Sequential Dependencies

The lack of support for the sequential dependencies hypotheses demanded further investigation. It appeared that two factors were influencing responses to the items: the format of the items and the number of relationships tested by the items (called information load). The structure and transfer items were ranked separately on both of these factors. The format rankings were based on the chance level of the items. For example, in a true-false pattern of six items, the probability of responding with the correct pattern would be $(1/2)^6$ or 1/64. These probabilities were then transformed to rank scores. The information load rankings were based on the number of structural relationships tested by each item. The number of relationships was taken from the test analysis (Appendix G). These numbers were also transformed to rank scores. It was expected that a high information load and low probability of chance success would be related to poor performance on the item. The actual difficulty level of each item was the proportion of Ss who received the maximum possible score on the item.

Item difficulties were chosen as the criterion because they seemed to be a good index of student performance and by ordering items on this basis a Guttman scale was produced. The reproducibility indices obtained by plotting item difficulties against Ss scores on S1, S2, T1, and T2, were generally high (Table 26).

Table 26

Guttman Reproducibility Coefficients on
Structure and Transfer

	S1	S2	T1	T2
O-NR	.83	.79	.87	.85
O-V	.81	.79	.89	.86
O-D	.82	.80	.87	.86
R-NR	.84	.81	.84	.82
R-V	.80	.82	.82	.81
R-D	.80	.83	.88	.82

A reproducibility (rep) index of .85 is usually considered the criterion for separating scales from non-scales (Torgerson, 1963). Although some rep indices were below .85 none of them were below .79. Thus the items generally met the criterion of scaleability.

The correlation between format (Fo) and information load (In) rankings was .726 for the structure items ($n = 20$, significant at the .01 level) and was .747 for the transfer items ($n = 10$, significant at the .01 level). These correlations indicated that as the format of an item became more difficult the information load of the item also increased. Rank correlations between the actual item difficulty and Fo and In for S1, S2, T1, and T2, were relatively high (Table 27).

The consistently high relationships of Fo and In to item difficulty indicated that these factors might have had a greater influence upon student performance than knowledge acquired from the reliability passage. It also

Table 27

Rank Correlations: Actual Difficulty with Format
and Information on Structure and Transfer

	S1		S2	
	Fo	In	Fo	In
O-NR	.68**	.51*	.73**	.58**
O-V	.61**	.53*	.59**	.51*
O-D	.61**	.50*	.72**	.32
R-NR	.59**	.51*	.64**	.33
R-V	.71**	.59**	.64**	.63**
R-D	.69**	.44*	.72**	.63**

	T1		T2	
	Fo	In	Fo	In
O-NR	.93**	.87**	.80**	.95**
O-V	.92**	.79**	.93**	.86**
O-D	.92**	.86**	.91**	.86**
R-NR	.88**	.86**	.85**	.90**
R-V	.89**	.80**	.91**	.80**
R-D	.85**	.88**	.89**	.86**

** $p < .01$

* $p < .05$

explains, in part, why performance was low. In order to have adequately tested the hypotheses about sequential dependencies between content areas, the Fo and In factors should have been controlled. This could have been done by reconstructing items where the chance level was constant or by transforming the item scores to correct for the Fo and In factors. But it is questionable that such a transformation would change the treatment effects in the present study,

since additional evidence (achievement variable and questionnaire data) supported the idea that adequate comprehension was not obtained after one reading of the passage.

Relationships among Main Variables

Contrary to expectation, time was not correlated with achievement, structure, or transfer for any of the experimental treatments. In general the correlations were positive, in accord with the hypothesis, but were not significant. In light of the previous analysis I_n and F_o appeared to be more important in determining S_s ' performance.

Errors on the training program were related to performance on achievement, structure, and transfer, indicating that those individuals who made errors on the training program also performed poorly on the criterion tests. Although not anticipated, this relationship is congruent with what one might expect if there is a general factor underlying performance on different tasks.

Correlations of performance on initial testing and on retesting were generally high as expected, with the achievement correlations being the highest. The lower consistency of the structure and transfer scores was probably due to the confounding variables of information load and item format, which made the items more difficult and perhaps more subject to differential interpretation by S_s upon retesting. The correlation pattern also suggested that the

Ss' knowledge or cognitive structure was more clear, more stable for topics covered by the achievement items than by the structure and transfer items. This inference is consistent with the original formulation of the items. A test which thoroughly examines structural relationships and transfer based upon these relationships is more apt to reveal confusions within an individual's organization of knowledge than a test which is not so thorough.

The high correlations between achievement and structure were not expected, rather structure and transfer relationships had been predicted. It would appear that the format and information load on the structure and transfer items lowered the correlations between these two variables. The high relationship between achievement and structure might be explained in terms of similarity of content tested. Of the three tests the achievement and structure tests were most similar and thus individuals who grasped the content in one area would be apt to respond appropriately on both tests. However the transfer test examined logically related, but slightly different areas than the other two tests.

Another consistent finding was that when patterns existed between dependent variables, the pattern was not unique to a given treatment but rather was common to all treatments. In general the results for Group R were consistent with the results for Group O (thus the data was pooled). This lack of interaction between patterns and treatments also indicated that the treatments were ineffective within the time allowed. The versions of the reliability

passage were objectively different, e.g., the number and form of repetitions of the structure. However, the data indicated that Ss reacted similarly to each of them. In order for differences among the treatments to occur, more than one reading would be necessary.

Because of the results, it is not possible to say that the diagrams functioned as effective blueprints of the structure of the material. In fact, the data implies the contrary, that they were not used as blueprints and probably confused rather than aided the Ss. However they did not confuse the students to the extent that the Ss' performance was below that of the other Ss in the NR and V treatments.

Implications of the Study

Before discussing various implications of the experimental results, the importance of the diagramming approach to structure of knowledge and cognitive structure measurement will be considered. The experimental study itself was not specifically designed to investigate the usefulness of diagramming for curriculum planning. Nevertheless the basic approach appears to be fruitful and is perhaps one of the most important aspect of the entire study. The importance of clearly defining structure of knowledge has been discussed previously. Diagramming is one answer to this problem.

Diagramming also has potential for individualized instruction, in determining the content and objectives of a course as well as testing

individual's knowledge. Although constructing tests based on a structural analysis is difficult, such a procedure provides an alternative to present procedures including tests based on Bloom's Taxonomy (Bloom, et al, 1956). For example, the scoring procedure emphasizes dependencies between ideas, rather than ignoring possible correlations between items. Test constructors usually attempt to construct items which are independent rather than dependent. Such procedure seems to be contrary to what is known about how individuals store and organize information. Diagramming also provides an absolute rather than a relative criterion for evaluating student achievement.

The systematic analysis discussed in (Chapter VI) of the structural relationships which Ss understood indicates the inadequacy of most achievement tests for providing diagnostic information about the individual, since this analysis showed that Ss had great difficulty with certain types of relationships. Such diagnostic information is usually not provided from achievement tests because they are not constructed to yield such information. However, in order to understand most subject matter disciplines, students need to grasp relationships, such as the causal and definitional ones in the present study. Perhaps lack of such comprehension is one of the reasons students have trouble with subject matter.

Tests built upon a structural analysis might reveal certain aspects of cognitive processes, such as Piaget's investigations have done. In both cases there has been an attempt to look beyond an individual's knowledge of "topics."

Rather than saying an individual "knows science and mathematics," it can be said that he "understands classifications and causal relationships," independent of content. Of course, the difficulty Ss had with causal relationships in the present study was tied to specific content, and only an investigation of other content with causal relationships would provide generality to the findings. However, the results suggest that there is need for more intensive study of the connection between how individuals handle various types of relationships and their success in school.

The unexpected results of the experimental study suggest several avenues for future research. First, one reading of the reliability passage did not produce comprehension. In studies using rather complex and lengthy material more than one reading appears to be necessary. Second, large individual differences occurred on the criterion variables. However, few background variables correlated with these scores, making it very difficult to determine the factors related to performance on the tests. Visual imagery might have been one of these factors, yet no adequate imagery tests existed. Presently the variability within each experimental treatment remains unexplained. Third, the difficulty that Ss had using the diagrams suggests that presenting diagrams within the material upon original learning may hinder, or at least not aid, understanding. Perhaps diagrams should be presented at later stages within the learning process. Such an approach would be partially supported by Travers, Heath, and Cahen (1968). In examining preferences for verbal,

graphic, and symbolic modes of presentation (the graphic being most similar to diagrams), Travers found that although students and teachers both preferred the symbolic mode, the teachers' preference for the graphic mode was consistently higher than the students' preference. This finding suggests that diagrams might be most beneficial when an individual has some competence within a content area. Fourth, the stability of scores over the one-week retention period suggests a need for Ebbinghaus-type memory studies on meaningful material in order to provide parametric data on retention curves for such content.

Two other results suggest important methodological considerations for similar studies. Group pressures on time spent reading passages is one methodological factor. Although time did not correlate with the criterion variables in the present study, it might be an important correlate or criterion variable in other studies. The high and consistent correlations between item difficulties and item format and information load strongly suggest that performance on tests is determined by factors other than knowledge itself, and that these factors should be considered in the interpretation of achievement test performance.

Finally, the systematic structural analysis indicates some inadequacies with most learning theories. Although research on concept attainment has shown that certain types of relationships are more difficult than others (disjunctive versus conjunctive), a theoretical explanation of why this occurs

has not been given by the cognitive or behaviorist learning theorists. From the results of this study it would appear that such factors can be explored more carefully within both the behaviorist and cognitive positions. In behavioral terms, the following type of problem might be investigated: can the nature of the connection or relationships between verbal stimuli (concepts, words) explain why certain associations are more dominant or learned more quickly than others. In cognitive terms, the following question might be explored: does the nature of the relationship between potential subsuming concepts and new material determine the speed of learning and the strength of retention. If learning theories are to adequately explain the learning of subject matter, investigation of such questions would seem to be imperative.

LIST OF REFERENCES

LIST OF REFERENCES

- Archer, E. J. The psychological nature of concepts. In Klausmeier and Harris (Eds.), Analyses of concept learning, New York, Academic Press, 1966, 37-49.
- Asch, S. E. Studies of independence and conformity, a minority of one against an unanimous majority. Psychol. Monogr., 1956, 70, No. 9 Whole No. 416.
- Ausubel, D. P. The psychology of meaningful verbal learning. New York, Grune and Stratton, 1963.
- Ausubel, D. P. Early versus delayed review in meaningful learning. Psy. in the Schools, 1966, 3, 195-198.
- Ausubel, D. P. & Fitzgerald, D. The role of discriminability in meaningful verbal learning and retention. J. educ. Psy., 1961, 52, 266-274.
- Ausubel, D. P., Robbins, Lillian C., & Blake, E. Retroactive inhibition and facilitation in the learning of school materials. J. educ. Psy., 1957, 48, 334-343.
- Ausubel, D. P. & Youssef, M. The role of discriminability in meaningful parallel learning. J. educ. Psy., 1963, 54, 331-336.
- Ausubel, D. P. & Youssef, M. The effect of spaced repetition on meaningful retention. J. gen. Psy., 1965, 73, 147-150.
- Berlyne, D. E. Structure and direction in thinking. New York, Wiley, 1965.
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. Taxonomy of educational objectives: Cognitive domain. New York, David McKay, 1956.
- Bourne, L. E. Human conceptual behavior. Boston, Allyn & Bacon, 1966, 73-79.
- Bruner, J. S. The process of education. Cambridge, Harvard Univ. Press, 1963.

- Bruner, J. S. Some theorems on instruction illustrated with reference to mathematics. In Theories of learning and instruction, NSSE Yrbk., 1964, 306-335.
- Bruner, J. S. Notes on the plenary sessions, Appendix B. In Bruner (Ed.), Learning about learning, a conference report, Washington, US Govt. Printing Off., 1966a, 245-276.
- Bruner, J. S. Theorems for a theory of instruction. In Bruner (Ed.) Learning about learning, a conference report, Washington, US Govt. Printing Off., 1966b, 196-211.
- Bruner, J. S., Olver, Rose R., & Greenfield, Patricia M. Studies in cognitive growth. New York, Wiley, 1966.
- Buros, O. K. Sixth mental measurements yearbook. Highland Park, New Jersey, Gryphon Press, 1965.
- Buros, O. K. Fifth mental measurements yearbook. Highland Park, New Jersey, Gryphon Press, 1959.
- Christensen, D. M. & Stordahl, K. E. The effect of organizational aids on comprehension and retention. J. educ. Psy., 1955, 46, 65-74.
- English, H. B. & English, Ava C. A comprehensive dictionary of psychological and psychoanalytic terms. New York, McKay, 1958.
- Fitzgerald, D. & Ausubel, D. P. Cognitive versus effective factors in the learning and retention of controversial material. J. educ. Psy., 1963, 54, 73-84.
- Flavell, J. H. The developmental psychology of Jean Piaget. New York, Van Nostrand, 1963, 17-19, 164-236.
- Gagné, R. M. The acquisition of knowledge. Psy. Rev., 1963, 69, 355-365.
- Gagné, R. M. The learning of principles. In Klausmeier and Harris (Eds.), Analyses of concept learning, New York, Academic Press, 1966, 81-95.

- Gagné, R. M. & Bassler, O. C. Study of retention of some topics of elementary nonmetric geometry. J. educ. Psy., 1963, 54, 123-131.
- Gagné, R. M., Mayor, J. R., Garstens, Helen L. & Paradise, N. E. Factors in acquiring knowledge of a mathematical task. Psychol. Monogr., 1962, 76, Whole No. 526.
- Gagné, R. M. & Paradise, N. E. Abilities and learning sets in knowledge acquisition. Psychol. Monogr., 1961, 75, Whole No. 518.
- Ghiselli, E. E. Theory of psychological measurement. New York, McGraw-Hill, 1964.
- Goss, A. E. Acquisition and use of conceptual schemes. In Cofer (Ed.), Verbal learning and verbal behavior, New York, McGraw-Hill, 1961, 42-69.
- Harary, F., Norman, R. Z. & Cartwright, D. Structural models, an introduction to the theory of directed graphs. New York, Wiley, 1965.
- Hartmann, G. The field theory of learning and its educational consequences, In The psychology of learning, NSSE Yrbk., 1942, 165-214.
- Johnson, P. E. Some psychological aspects of subject-matter structure. J. educ. Psy., 1968, 58, 75-83.
- Johnson, T. J. A methodology for the analysis of cognitive structure. Paper presented at the meeting of the American Educational Research Association, Chicago, February 1968.
- Lovell, K. Educational psychology and children. London, Univ. London Press, 1964, 96-99.
- McKellar, P. Imagination and thinking. London, Cohen & West, 1957, 19-31, 51-72.
- Malter, M. J. Children's ability to read diagrammatic materials. Elem. Sch. J., 1948, 49, 98-102.
- Merrill, M. D. & Stolurow, L. M. Hierarchical preview versus problem oriented review in learning an imaginary science. Am Educ. Res. J., 1966, 3, 251-261.

- Miller, G. A. The magical number seven, plus or minus two. Psy. Rev., 1956, 63, 81-97.
- Morrissett, I. The new social science curricula. In Morrissett (Ed.), Concepts and structure in the new social science curricula, New York, Holt Rinehart & Winston, 1967, 3-10.
- Newton, J. M. & Hickey, A. E. Sequence effects in programmed learning of a verbal concept. J. educ. Psy., 1965, 56, 140-147.
- Novak, J. D. The role of concepts in science teaching. In Klausmeier & Harris (Eds.), Analyses of concept learning. New York, Academic Press, 1966, 239-254.
- Posner, M. I. Memory and thought in human intellectual performance. Brit. J. Psy., 1965, 56, 197-215.
- Reitman, W. R. Cognition and thought, an information processing approach. New York, Wiley, 1965.
- Reynolds, J. H. & Glaser, R. Effects of repetition and spaced review upon retention of a complex learning task. J. educ. Psy., 1964, 55, 297-308.
- Senesh, L. Organizing a curriculum around social science concepts. In Morrissett (Ed.), Concepts and structure in the new social science curricula, New York, Holt Rinehart & Winston, 1967, 21-38.
- Scott, W. A. Cognitive complexity and cognitive flexibility. Sociometry, 1962, 25, 404-414.
- Sheffield, F. Theoretical consequences in the learning of complex sequential tasks from demonstration and practice. In Lumsdaine (Ed.), Student response in programmed instruction, Washington, Natl. Acad. of Sciences, Natl. Res. Council, 1961, 13-32.
- Sheffield, F. D., Margolius, G. J. & Hoehn, A. J. Experiments on perceptual mediation in the learning of organizable sequences. In Lumsdaine (Ed.), Student response in programmed instruction, Washington, Natl. Acad. of Sciences, Natl. Res. Council, 1961, 107-116.
- Smith, K. U. & Smith, Margaret F. Cybernetic principles of learning and educational design. New York, Holt Rinehart & Winston, 1966, 329-352.

- Torgerson, W. S. Theory and methods of scaling, New York, Wiley, 1963.
- Travers, K. J., Heath, R. W. & Cohen, L. S. Cognitive preferences in mathematics. Paper presented at Annual Meeting of the American Educational Research Association, Chicago, Illinois, February, 1968.
- Woodworth, R. S. Experimental psychology, New York, Henry Holt, 1938, 39-47.
- Vernon, M. D. The instruction of children by pictorial illustration. Brit. J. educ. Psy., 1953a, 24, 171-179.
- Vernon, M. D. The use and value of graphic material within a written text. Occupational Psy., 1952, 26, 96-100.
- Vernon, M. D. Presenting information in diagrams. AV Comm. Rev., 1953b, 1, 147-158.
- Zajonc, R. B. The process of cognitive tuning in communication. J. Abn. & Social Psy., 1960, 61, 159-167.

APPENDICES

APPENDIX A
PILOT QUESTIONNAIRE

Diagram Treatment

A. Interpretation of Diagrams

Go back through the material on the interpretation of diagrams and cross out any sections which are badly written (not clear). Did you have trouble remembering to mark your answer in this training program? Yes No.

If so, would additional reminders to indicate your response help?

Yes No

B. Reliability Passage

Go back through the reliability passage and cross out any sections which are badly written.

Did you examine the diagrams while reading? Yes No

If so, did you have any problem interpreting them in the reliability passage? Yes No

Did the review diagram present any problems? Yes No

Did you know where to start? Yes No

Did you understand the interconnections between diagrams?

Yes No

Did you understand each of the sub-diagrams? Yes No

Additional comments:

Did the diagrams hinder your understanding of the material?

Yes No

If so, would you have preferred a verbal statement instead?

Yes No

In general, was the passage difficult to understand? Yes No

C. Test

Go back through the test and cross out any badly written items.

Were the instructions for each item clear? Yes No

If not, which instructions were not clear?

Did any items cue off the answer to another question? Yes No

If so, which ones?

In general did the items seem

- a. difficult
- b. some easy, some hard
- c. easy

Was there any particular type of item which seemed particularly difficult?

Did you use the diagrams in any particular way when reading the passage and/or taking the test? Yes No

If so, please briefly describe this process.

Passage:

Test:

If the diagrams helped you, did the small diagrams or the review diagram help you the most?

Small

Review

No difference

Verbal Treatment

A. Reliability Passage

Go back through the reliability passage and cross out any sections which are not clear .

Did the reliability passage seem disconnected in places? Yes No .

If so, mark these places with a check .

In general , was the passage difficult to understand? Yes No

B. Test

Go back through the test and cross out any badly written items .

Were the instructions for each item clear? Yes No

If not , which instructions were not clear?

Did any items cue off the answer to another question? Yes No

If so, which ones?

In general , did the items seem

- a. difficult
- b. some easy , some hard
- c. easy

Was there any particular type of item which seemed particularly difficult ?

Did you notice the small review passages placed within the text as well as at the end? Yes No

If so, did these help you while reading and/or taking the test ?

Yes No

If the passages helped you, did the small passages or large review help you the most?

Small

Review

No difference

No-Review Treatment

A. Reliability Passage

Go back through the reliability passage and cross out any sections which are not clear.

Did you use any particular process while reading the material in order to understand it?

In general, was the passage difficult to understand? Yes No

B. Test

Go back through the test and cross out any badly written items.

Were the instructions for each item clear? Yes No

If not, which instructions were not clear?

Did any items cue off the answer to another question? Yes No

If so, which ones?

In general, did the items seem

- a. difficult
- b. some easy, some hard
- c. easy

Was there any particular type of item which seemed particularly difficult?

How did you recall the material when taking the test?

APPENDIX B
TREATMENT QUESTIONNAIRE

Questions Common to All Treatments

1. Was the content of the reliability passage (in general) new to you?
Yes No
2. Did you enjoy reading the reliability passage? Yes No
3. Which parts of the reliability passage were difficult to understand?

Mark the ones which apply:

- ☐ Importance of reliable tests
- ☐ Distinction between systematic and unsystematic factors
- ☐ Distinction between constant unsystematic and varying unsystematic factors
- ☐ Methods of estimating reliability coefficients
- ☐ Concepts of reliability coefficient and correlation coefficient
- ☐ Parallel forms of a test versus parallel tests

Additional Questions - Diagram Treatment

1. Did you examine the small diagrams presented within the reliability passage? Yes No

If you answered "yes" to the above question, did you have trouble interpreting these small diagrams within the passage?

Yes No

If so, why did you have problems?

2. Did you examine the large review diagram at the end of the reliability passage? Yes No

If you answered "yes" to the above question:

- a. Did you have trouble interpreting the six sub-diagrams?
Yes No

If so, why?

b. Did you examine the interconnections between the sub-
diagrams? Yes No

If so, did you examine the connections systematically OR in a
more or less random, non-orderly fashion? (underline the
appropriate description)

3. How did you use the small diagrams when you read the passage?

(Mark the statements which apply)

- _____ As a repetition of the previous material
- _____ To integrate the previous material
- _____ As a check on what you had previously learned (read)
- _____ To organize the material
- _____ As a way to remember the material in a spatially organized (visual)
form, rather than or in addition to a verbal form

Other use(s) of diagrams _____.

4. How did you use the diagrams when answering the test items (this applies
to both the first and second testing periods)?

Mark the ones which apply.

- _____ Visualized the related diagram to an item and "read off" the answer
from it
- _____ Instantly recognized that a diagram had dealt with the topic covered
by an item but did not visualize the diagram
- _____ Vaguely remembered that a diagram had dealt with the topic covered
by an item
- _____ Did not recall while answering any of the questions that diagrams could
have been specifically related to the items

Other use(s) of diagrams _____

Additional Questions - Verbal Treatment

1. Did you examine the small review passages presented within the reliability passage? Yes No
2. Did you read the large review section at the end of the reliability passage? Yes No
3. How did you use the review passages when you read the passage?

(Mark the statements which apply .)

- _____ As a repetition of the previous material
- _____ To integrate the previous material
- _____ As a check on what you had previously learned (read)
- _____ To organize the material
- _____ As a way to remember the material in a verbal form

Other use(s) _____

4. How did you use the review passages when answering the test items?

(This applies to both the first and second testing periods.) Mark the ones which apply .

- _____ Instantly recognized that a review passage had dealt with the topic covered by an item
- _____ Vaguely remembered that a review passage had dealt with the topic covered by an item
- _____ Did not recall while answering any of the questions that a review passage could have been related specifically to the items

Other use(s) _____

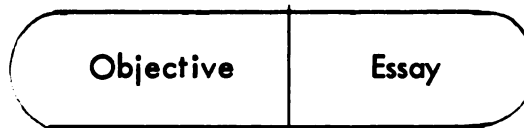
APPENDIX C
DIAGRAM INTERPRETATION
TRAINING PROGRAM

Suppose we read the following passage .

Most achievement tests can be included in one of two categories, objective or essay . With an essay test a student is required to plan his own answer and express it in his own words, whereas an objective test requires him to choose among several designated alternatives .

Through the use of a Venn diagram we can easily represent this idea of two types of achievement tests .

Types of Achievement Tests



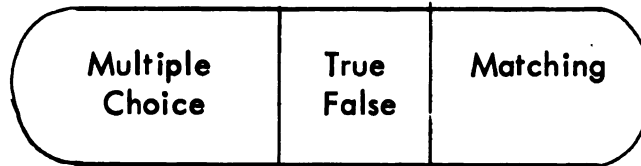
The ellipse represents the entire class of achievement tests, i.e., all possible achievement tests. It was then divided or partitioned into two parts (of course, it could have been divided into more than two parts) representing the two types of achievement tests, objective and essay . With this type of diagram we have illustrated the idea of classification (smaller categories grouped under larger ones) .

Please turn to the next page .

Page 3

Suppose we had the following Venn diagram.

Objective Tests



Using this diagram we could then say that (choose one alternative)

a. All multiple choice tests are objective.

b. All objective tests are true-false.

If you chose alternative a turn to page 5.

If you chose alternative b turn to page 4.

Please remember to mark your choice throughout the passage.

Page 4

No, a true-false test is only one type of objective test.

The diagram shows that the category of objective tests is rather broad, being partitioned or split into three types of tests; true-false, matching, and multiple-choice. Therefore not all objective tests are true-false ones. The three types of objective tests are nonoverlapping or distinct from one another as indicated by the partition lines (refer back to page 3 if necessary).

Now go on to page 6.

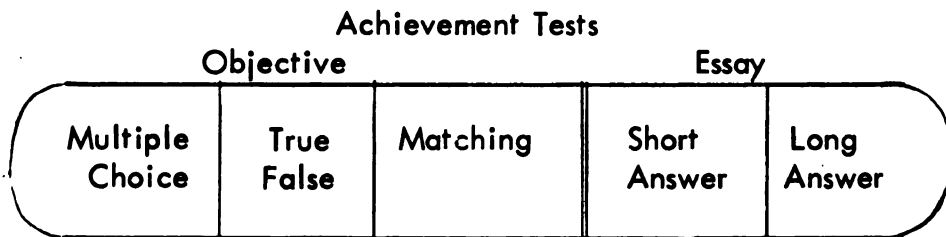
Page 5

Yes, the Venn diagram represents the idea that we can classify objective tests into three different types; multiple-choice, true-false, and matching. These types are nonoverlapping or distinct from one another as indicated by the partition lines (refer back to page 3 if necessary).

Now turn to page 6.

Page 6

Now examine this diagram.



Using the above diagram it would be correct to say that

- At the lowest level there are five types of achievement tests, three of which are essay and two of which are objective.
- If an individual wanted an essay achievement test, you could give him a short answer test.

If chose a turn to page 7.

If chose b turn to page 8.

Page 7

It is correct that at the lowest level of achievement tests there are

five types, but these are grouped differently; three are objective (multiple-choice, true-false, matching) and two are essay (short and long answer) not vice versa. The darker line in the center of the ellipse separates objective from essay tests. In turn these two partitions or parts are then divided further by the use of lighter lines within each part. (Refer back to page 6 if necessary).

Now proceed to page 9.

Page 8

Yes, the diagram shows that one of the two types of essay tests is short answer, the other being long answer. The broad category is first split into objective and essay, represented by the darker line in the center of the ellipse. Each of these types is then divided further, symbolized by the lighter lines within each of these larger parts - three types of objective tests and two types of essay tests. (Refer back to page 6 if necessary).

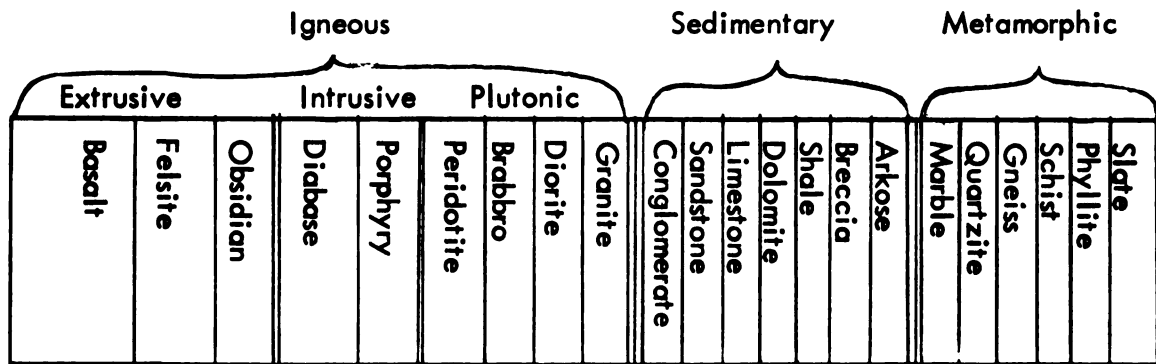
Now turn to page 9.

Page 9

This basic idea of classification and separation into smaller categories could be extended and varied indefinitely and is not limited to the exact examples given here. For instance, consider the classification of rocks.

Page 9 - continued

Rocks



Here there are three great classifications of rocks; igneous, sedimentary, and metamorphic, with igneous then being split into three other basic types. Finally at the finest level of classification there are 22 classes; six within metamorphic, seven within sedimentary, and nine within igneous.

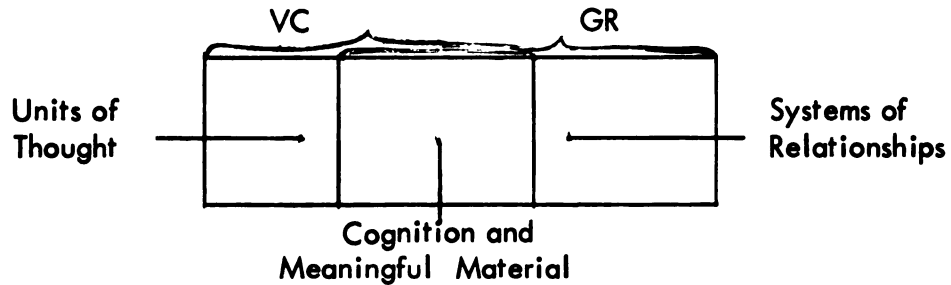
Go ahead to the next page.

Page 10

Venn diagrams can also be used to indicate overlap among concepts. Suppose we look at some factors which are often a part of intelligence tests; verbal comprehension, general reasoning, and spatial orientation. We could define verbal comprehension as including cognition, meaningful material and units of thought; general reasoning involving cognition, meaningful material and systems of relationships; spatial orientation as involving cognition, figures, and systems of relationships.

Page 10 - continued

The relationships between verbal comprehension and general reasoning can be represented as follows:



Verbal comprehension and general reasoning overlap because they both involve

- a. cognition and meaningful material
- b. systems of relationships

If you answered a, turn to page 11.

If you answered b, turn to page 12.

Page 11

Yes, the brackets indicating verbal comprehension and general reasoning overlap in the rectangle marked cognition and meaningful material. The diagram also represents information about the other two areas - units of thought is not a part of general reasoning and systems of relationships is not a part of verbal comprehension. (Turn back to page 10 if desired).

Now turn to page 13.

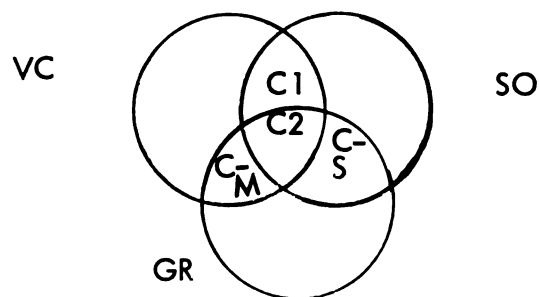
Page 12

No, the brackets indicating verbal comprehension and general reasoning overlap in the rectangle marked cognition and meaningful material, not the one marked systems of relationships. In fact, systems of relationships is not a part of verbal comprehension, while units of thought is not a part of general reasoning. (If desired refer back to page 10).

Turn to page 13.

Page 13

Letting circles instead of rectangles (the type of shape used in a Venn diagram is not crucial to understanding the material represented by it) stand for these aspects of intelligence, we can represent the relationship among all three as follows:



We can then say that all three overlap in the area of

- a. cognition 2 (C2)
- b. cognition and systems (C-S)

If chose a turn to page 14.

If chose b turn to page 15.

Page 14

Yes, all three circles intersect in this area. Note that C1 represents the overlap between verbal comprehension and spatial orientation, while C-S (cognition-systems) represents the overlap between general reasoning and spatial orientation. Refer back to page 13 if necessary.

Now turn to page 16.

Page 15

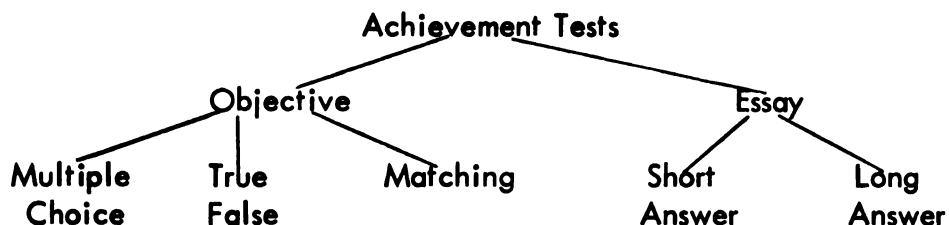
No, C-S represents the overlap between general reasoning and spatial orientation only (only two circles intersect in this area). Instead C2 represents the overlap of all three - verbal comprehension, spatial orientation and general reasoning (all three circles intersect here). Note also that C1 represents the overlap of verbal comprehension and spatial orientation. Refer back to page 13 if necessary.

Now please turn to page 16.

Page 16

Other types of diagrams could represent the same material. Both the rock and test information could be represented by a "tree" graph. For example with the achievement tests we could have

Page 16 - continued



Here the lines represent the breakdown of a large class into smaller ones.

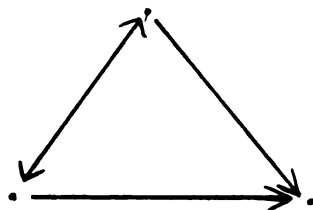
This division occurs at each level of classification, proceeding from a broad, inclusive category to narrower, less inclusive ones.

In other terms, the names of tests can be represented by points with lines between these points representing how these tests are related to the classification of tests at the next higher and/or lower level. So here we have points with lines between them; the lines arranged in a certain manner.

Go to the next page.

Page 17

The basic idea of points with lines connecting them is what we might generally call a line graph, e.g.,



It can be used to represent many different things such as flow charts in chemistry, time lines in history, cause-effect sequences, etc. The preceding tree graph

Page 17 - continued

illustrating a way of classifying tests would be called a type of line graph.

Since you can define both points and lines as you wish, i.e., have them represent a variety of concepts and relationships, this basic type of diagram is quite flexible. The following examples illustrate some ways in which it can be used.

Turn to the next page.

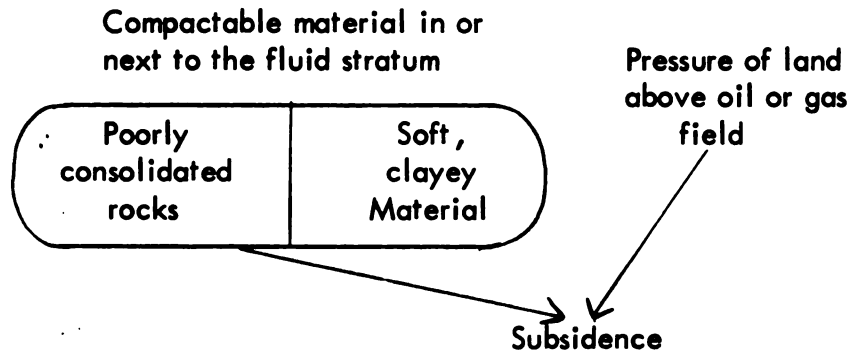
Page 18

Consider the topic of geological subsidence or sinking of lands.

Geological subsidence or sinking of lands results from tapping the earth for oil or gas. Near Long Beach, California the land above the Wilmington oil field sank until it had become a bowl up to 26 feet deep over an area of 22 square miles. The slow subsidence of the land ruined buildings, cracked pavements, twisted railroad tracks, wrecked bridges, sheared off oil wells, and did extensive damage to a power plant and the Long Beach Naval Shipyard resulting in total damage of about \$100 million.

The explanation for such phenomenon is as follows. Liquid or gas is generally drawn from a stratum of porous rock whose pores are filled with the fluid under pressure. If the rock is well consolidated (if its grains are well cemented together) it will usually continue to support the weight of the rock and earth on top after the fluid is withdrawn. However, if the fluid-holding rock is a poorly consolidated, easily-molded sandstone, once the supporting pressure of the fluid has been withdrawn from its pores the pressure of the overburden compacts the rock, and the ground above subsides by the amount by which the rock is compressed. Other factors besides the mechanical strength of the fluid-containing rock may contribute to subsidence. For example, subsidence is more likely if soft, clayey material (which is easily compacted) is present in or next to the fluid stratum.

One way to diagram the relationship between subsidence and its causal factors is as follows, where "↓" represents the concept of cause.



True or False

We can say then that pressure of the overlying land is the only reason for the occurrence of subsidence.

If answered true, turn to page 19.

If answered false, turn to page 20.

(Please mark your answer to each question).

No, the two arrows pointing to the word "subsidence" indicate two major factors, not one, leading to geological subsidence. Just pressure from the land above is not enough. Compactable material is therefore the other factor. Note that in this big diagram the classification of compactable materials into two types, clay and poorly consolidated rocks is represented by a Venn diagram. Thus either type can contribute to subsidence. (Refer back to page 18). Now turn to page 21.

Page 20

Correct. The two arrows pointing to the word "subsidence" indicate two major factors, not one. Note, that in this diagram the classification of compactable materials into two types, clay and poorly consolidated rocks, is represented by a Venn diagram. Thus either type can contribute to subsidence; although we would say that in general there are two major factors causing subsidence, compactable materials and pressure. (Refer back to page 18 if necessary).

Please turn to page 21.

Page 21

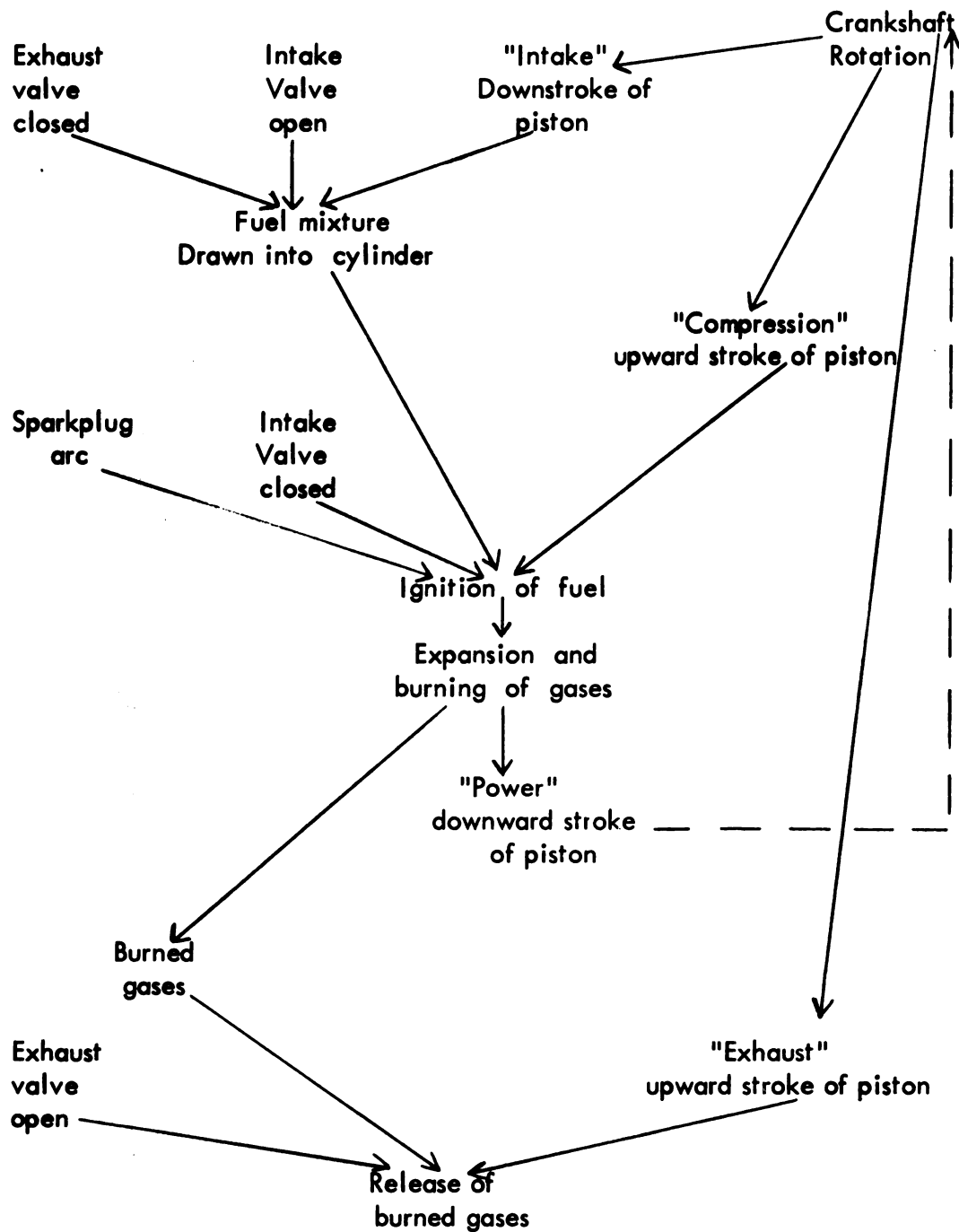
Let us consider another situation - the various piston strokes (intake, compression, power, exhaust) in a gasoline engine. The sequence of events can be briefly described.

The "intake" stroke is downward with the intake valve open to let the gasoline-air mixture into the cylinder. On the upward "compression" stroke the fuel mixture is compressed. The sparkplug ignites the fuel when the piston head is at the top of the compression stroke. The expanding gases then deliver a "power" stroke pushing the piston downward. The upward "exhaust" stroke pushes the burned gases out the exhaust valve. The crankshaft, changes the up and down (reciprocating) motion of the piston to turning (rotary) motion and vice versa.

We could diagram these cause-effect relationships as indicated on the next

page, where " ↓ " essentially represents "causes."

Page 22



We can then say that the rotation of the crankshaft has a direct part in causing every type of piston stroke except the

Page 22 - continued

a. compression stroke

b. power stroke

If answered a turn to page 24.

If answered b turn to page 23.

Page 23

Yes, this is correct because there is no arrow from the words "crankshaft rotation" to "power stroke" indicating that the direct cause for this stroke is elsewhere. (Refer back to page 22 if necessary.)

Turn to page 25.

Page 24

No, the direct solid arrow from the words "crankshaft rotation" to "compression stroke" indicates that rotation of the crankshaft serves to push the cylinder upward, in this case the compression stroke. There is no line (arrow) to the words "power stroke," indicating another cause for this stroke. (Refer back to page 22 if necessary).

Turn to page 25.

Page 25

True-False question

We can say that the intake of the fuel mixture into the cylinder chamber was due to the open intake valve only. Refer back to page 22.

If answered true, turn to page 26.

If answered false, turn to page 27.

Page 26

No, the multiple number of arrows or lines focusing on "Fuel mixture drawn into cylinder" means that there are several joint reasons, not just one, for this event.

Now turn to page 28.

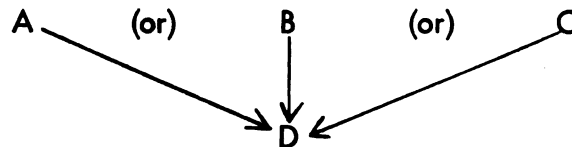
Page 27

Yes, you have correctly interpreted the multiple number of arrows (lines) focusing on "Fuel mixture drawn into cylinder" as representing several joint reasons, not just one, for this event.

Now turn to page 28.

Page 28

In the diagram representing the piston strokes of a gasoline engine, several arrows from different points converging on one point meant that all these conditions were necessary to cause this one event, not just one condition alone. If we had a situation where several independent conditions could cause the same event, i.e., each cause it without the others being present, we would probably diagram this situation as follows:



Here A or B or C alone could cause D as indicated by the word "or" in parentheses.

Please turn to the next page.

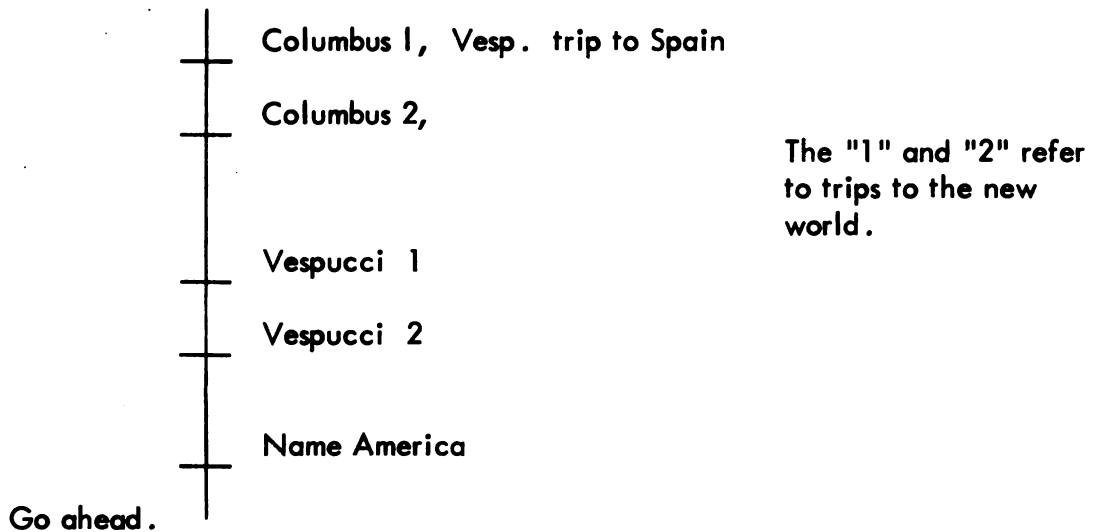
Page 29

Suppose we look at this same type of diagram in the context of historical material, for instance the naming of North and South America after Amerigo Vespucci rather than Columbus. We might present in a written passage the position that Columbus believed he had found India or China while Vespucci was the first to realize that the body of land was really a new continent. Various events in Vespucci's life led up to his exploration of the new

Page 29 - continued

world, including an interest in and a study of geography and his conviction that there was a southern route to India if you only went far enough south. Several sequences will be examined here (a) a timeline of important events, (b) name of the new world, (c) beliefs of the location of the water route to India and (d) why the new world was named after Vespucci.

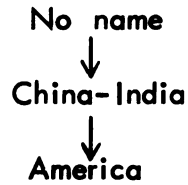
This diagram is simply a time line of events. The relative distance between the horizontal lines represents the ordering of events, spaced on the vertical line which represents the continuum of time.



Page 30

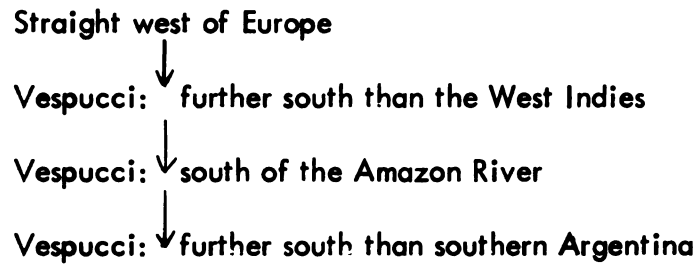
If we examine a sequence of names for the new world we find

Page 30 - continued



Here we can simply let the lines " —————> " stand for the word "then" or "to," i.e., at first it had no name then China-India and then America.

A similar interpretation applies to the sequence dealing with people's beliefs of a location of a water route to India.



With this diagram we could then say that the water route to India

- a. was believed to be below southern Argentina and later south of the Amazon River
- b. was believed to be south of the West Indies then south of southern Argentina

If answered a, then turn to page 31.

If answered b, then turn to page 32.

Page 31

You have apparently confused the order of events. The line or arrow " $\downarrow$ " represents the word "then." Thus if we have the sequence, $A \longrightarrow B$, it means that A is followed by B, not precedes B. Therefore in this diagram the direction of the arrow means that belief in the south of southern Argentina route followed the south of Amazon River belief. (Refer back to page 30 if desired).

Please turn to page 33.

Page 32

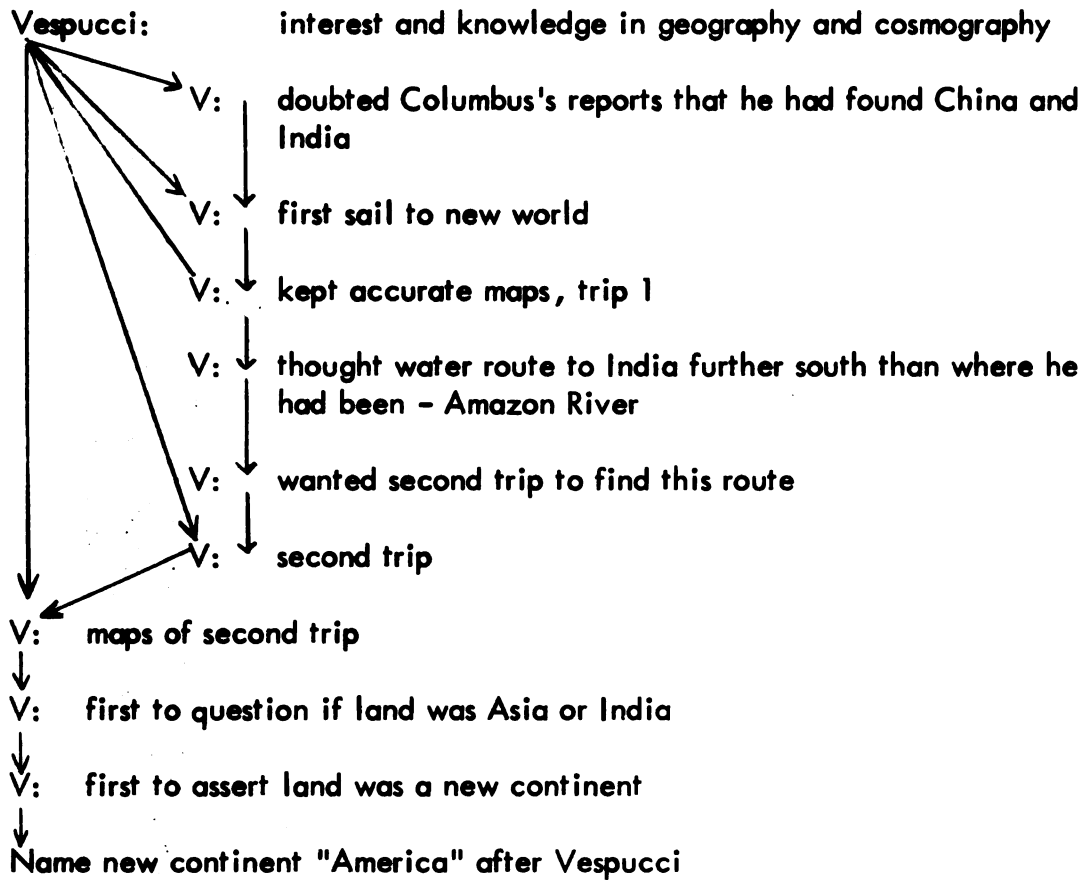
You have interpreted the diagram correctly. The line or arrow " $\downarrow$ " represents the word "then." Thus when we have the sequence $A \longrightarrow B$, it means that A is followed by B, not precedes B. This is the situation concerning the belief in a route south of the West Indies and the belief in it further south than southern Argentina (being respectively "A" and "B".) (Refer back to page 30 if desired.)

Please turn to page 33.

Page 33

The last sequence is that of why the new world was named after Vespucci.

Page 33 - continued



From this diagram information we could say that America was named after Amerigo Vespucci because

- a. Amerigo Vespucci made very accurate and detailed maps of new land on his second trip
- b. Vespucci was the first to assert that the land Columbus had found was a new continent.

If answered a, turn to page 35.

If answered b, turn to page 34.

Page 34

Yes, an arrow leading directly from one event to another implies a causal relationship. This is not the case if another event intervenes between the two and there is no arrow directly connecting the first with the last event. (Refer back to page 33 if desired).

Now please turn to page 36.

Page 35

No, because of these accurate and detailed maps he began to question if this body of land was a new continent. This is indicated by the arrow directly relating to these two events. (Refer back to page 33 if desired). The reason for the naming was because Vespucci was the first to assert that this was a new continent (represented by the arrow directly relating to these two events). If an event intervenes between two others then the first event does not cause the last one unless there is a direct arrow connecting the two.

Now please turn to page 36.

Page 36

Finally, we can interconnect all of these diagrams on the basis of time - see page 37. Here the same diagrams are reproduced with dashed lines drawn among them to provide a direct connection with the time line.

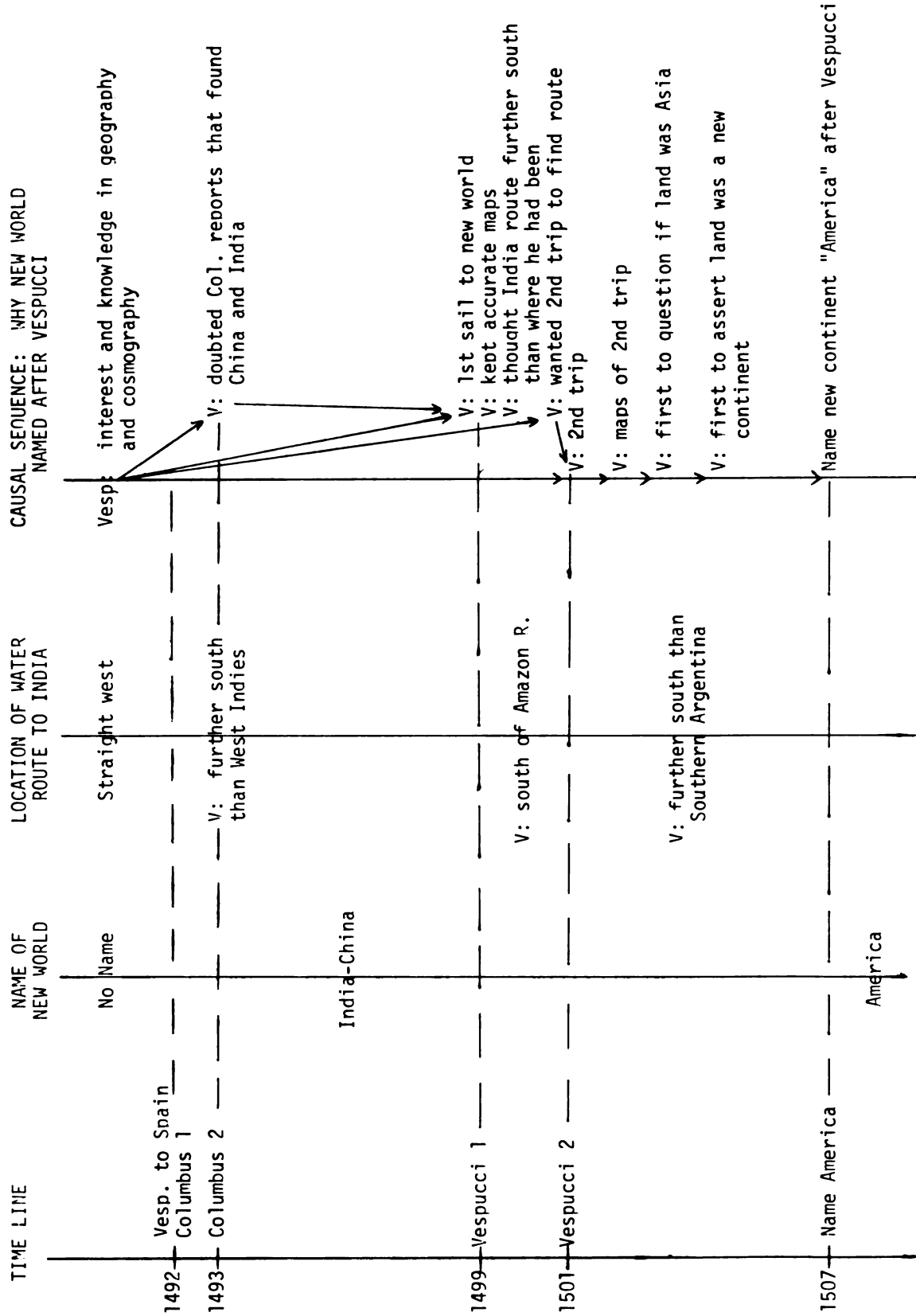
Page 36 - continued

From considering this more complex diagram we can say that

- a. Columbus's first trip and Vespucci's belief that the southern route to India was south of the Amazon River coincided.
- b. Vespucci's belief that the route to India was south of southern Argentina occurred on his second trip.

If you answered alternative a, turn to page 38.

If you answered alternative b, turn to page 39.



Page 38

No, Columbus's trip was before. The cross-connecting dashed lines as well as the area between them indicate events on each separate sequence which occurred at approximately the same time. Following the line directly across from "Columbus 1" (on first sequence) to "location of water route to India" sequence we find that before Columbus's first trip the water route was believed to be straight west of Europe, but after the trip Vespucci concluded that it was south of the West Indies, and only even later that he believed it was south of the Amazon River. (Refer back to page 37 if necessary).

Proceed to page 40.

Page 39

Yea! The dashed lines and the area between them indicate events on each separate sequence which occurred at approximately the same time. You correctly related the events on the "location of water route to India" sequence and the "causal sequence." (Refer back to page 37 if desired).

Proceed to page 40.

Page 40

The preceding examples have illustrated only a few variations of a line graph. Other concepts such as "equivalence," "implication," "greater

Page 40 - continued

than," etc. could all be represented by such lines. We will now look at one more type of diagram.

This type of diagram is what might be labelled a table. It is quite appropriate for showing descriptive relationships, characteristics of certain objects, etc.

Consider the official languages of the countries in the Western world. ("Country" refers here to even a nation which is not independent). We might divide the countries into the fairly large categories of North, South, and Central America. A table of the countries and corresponding languages could then be constructed with the columns being the languages and the rows being the countries or vice versa. The result might look something like that on the next page. The "X" essentially represents the fact that a certain language is the official language of a certain country.

		Spanish	English	French	Portuguese	Dutch
North America	United States		x			
	Canada		x	x		
	Mexico	x				
Central America	Panama	x				
	Brit. Honduras		x			
	Guatemala	x				
	Honduras	x				
	Nicaragua	x				
	Costa Rica	x				
	El Salvador	x				
	Cuba	x				
	Dominican Rep.	x				
	Haiti			x		
	Jamaica		x			
	Trinidad & Tobago		x			
	Puerto Rico	x	x			
South America	Argentina	x				
	Brazil				x	
	Chile	x				
	Uruguay	x				
	Paraguay	x				
	Peru	x				
	Ecuador	x				
	Bolivia	x				
	Columbia	x				
	Venezuela	x				
	Brit. Guiana		x			
	Surinam					x
	French Guiana			x		

Page 41 - continued**True-False question**

The dominant language in Central and South America is Spanish while in North America it is English.

If answered true, turn to page 43.

If answered false, turn to page 42.

(Please indicate your answer to each question).

Page 42

No, note that of the three countries in North America two of them have English as an official language while in South and Central America practically all of the countries list Spanish as the official language. This is indicated by the checks (X) for a language in correspondence with the classification of the countries. (Refer back to page 41 if necessary).

Proceed to page 44.

Page 43

Yes, this is indicated simply by the number of checks for any language while at the same time considering the classification of countries that is given.

Proceed to page 44.

Page 44**True-False question**

No country has more than one official language. (Turn back to page 41 for the table).

If answered true, turn to page 45.

If answered false, turn to page 46.

Page 45

This is incorrect. To determine this simply check each row, i.e., each country, to see if one or more than one check appears. If this is done it can be seen that Canada and Puerto Rico each have two official languages. (See page 41 if necessary).

Turn to page 47.

Page 46

Yes, Canada and Puerto Rico each have two official languages. As you know, all that is required to determine this is to count the number of checks for each row, i.e., each country.

Go to page 47.

Page 47

Actually such a table reflects the historical heritage of many of these countries. If this diagram were related to other historical events, also in diagram form, we could have a more complete picture of the reasons for the official language(s) of each country.

We're almost done. Proceed to the next page.

Page 48

As a last example of the table form of diagram consider the types of rocks referred to before: igneous, sedimentary and metamorphic. Each can be described in a very broad way as a product of a formation process. The following table summarizes this information.

Types of Rocks			
Igneous	Sedimentary	Metamorphic	
X			Molten rock which has cooled and hardened
	X		Rock grains locked together by pressure and cementing material
		X	Rock with changed mineral content

The check (X) represents the fact that a given type of rock

Page 48 - continued

a. can be defined in terms of certain characteristics

b. causes certain characteristics

If you answered a, please turn to page 50.

If you answered b, please turn to page 49.

Page 49

No. The check represents the fact that each type of rock is defined by a certain characteristic, these being the end product of specific formation processes (not specified here). A rock type does not cause certain characteristics, but rather is defined in terms of them.

Turn to page 51.

Page 50

Correct. The check essentially is representing a descriptive or defining relationship. e.g., metamorphic rocks are ones with changed mineral content. The reason for each type of rock having certain characteristics is that it is caused by a formation process (not described here) which has resulted in a certain end product.

Turn to page 51.

Page 51

Only one check in each column means that the definitions of the rock types (refer back to page 48 if necessary) .

a. overlap

b. do not overlap

If a, turn to page 52 .

If b, turn to page 53 .

Page 52

No, the definitions do not overlap. If they did this fact would be represented by more than one check in each column. In most cases the pattern of checks is quite important in interpreting tables; both the presence and omission of a check having implications for the subject matter at hand.

Go to page 54 .

Page 53

Yes, if the definitions did overlap then there would be more than one check in each column. In most cases the pattern of checks is quite important in interpreting a table; both the presence and omission of a check having implications for the subject matter at hand.

Go to page 54 .

Page 54

You have now concluded a program on the nature and interpretation of various diagrams which can be used to clarify and represent ideas presented in written material. Not all variations of these three basic types of diagrams have been presented, but these other variations would simply be extensions of the basic principles already presented. Of course, small diagrams can be connected to other small ones resulting in a rather large more complex diagram form. In general, the diagrams presented here have been rather simple in structure.

Thank you.

APPENDIX D
OUTLINE OF RELIABILITY PASSAGE

- I . General definition of reliability
 - A. Reliability applied to education
- II. Importance of reliability in testing
 - A. Differences among individuals, same test
 - B. Assignment of individuals to groups
 - C. Prediction
 - D. Differences among traits of an individual
- III. Systematic variation and unsystematic variation
- IV. Systematic factors and unsystematic factors
- V. More precise definition of reliability
 - A. Parallel tests
 - B. Correlation coefficient
 - C. Reliability coefficient
 - D. Relationship between correlation and reliability coefficient
- VI. Unsystematic factors
 - A. Varying unsystematic factors
 - B. Constant unsystematic factors
 - C. Comparison of varying and constant unsystematic factors
 - D. Comparison of constant unsystematic and systematic factors
- VII. Methods of estimating reliability coefficients
 - A. Test-retest
 - B. Parallel forms
 - C. Internal consistency

APPENDIX E
RELIABILITY PASSAGE

One of the important aspects of any measuring instrument is its reliability. Reliability of measurement refers to the consistency with which an instrument measures whatever it purports to measure. Obviously all measuring instruments are not perfectly accurate or consistent. Error is unavoidably involved in any measurement, but the goal of measurement specialists is to reduce these errors to a minimum. To the extent that these instruments deviate from yielding perfectly consistent measurements, i.e., their scores vary unsystematically, they are said to be unreliable. Thus the measurements from a wooden ruler are apt to be more reliable than those from a rubber one, since the latter measuring instrument fluctuates with the temperature, tension, etc., yielding inconsistent results. D1

In the field of education, reliability usually refers to the consistency with which a test measures whatever it purports to measure. Generally this consistency reflects the degree to which the test may be considered stable or may be depended upon to yield similar test results under similar circumstances. Tests may be achievement, aptitude, or personality measures. These tests give us quantitative descriptions of individuals (a score) in terms of the extent to which individuals possess or manifest various traits or abilities. For instance a high score means that an individual possesses more of a certain trait (for example, happiness) than an individual who has a lower score. Ordinarily we are interested in these quantitative descriptions or scores because of their usefulness in permitting us to make comparisons among individuals on a given

trait and within individuals on different traits, for predicting other types of behavior, and for evaluating the effects of various factors upon an individual's performance. As a consequence when we measure an individual, we hope to obtain a score that will give us a precise characterization of him.

If we administer the same test several times to an individual, we may observe unsystematic variation or little self-consistency in his scores. For example, if we give a psychological test to an individual on several different occasions, he might obtain scores of 83, 65, 75, 89, 80. The degree of self-consistency among the scores earned by an individual is termed reliability of measurement or simply reliability. When scores are not self-consistent it means that we cannot depend too much upon any single score earned by an individual since on another application of the same test he might earn quite a different score.

Unreliable scores are of little value when we wish (a) to compare two or more individuals on the same test, (b) to assign individuals to groups or classes, (c) to predict other types of behavior, or (d) to compare different traits or abilities of an individual. Let us consider examples of these common uses of tests and the importance of reliability in these situations.

The extent to which we are willing to trust the difference between the scores earned by two individuals on a test as reflecting a real or stable difference between them in the trait being measured by the test is a function of the reliability of that test. Sometimes we wish to know whether one person

is superior to another in the traits or abilities measured by a particular test .

If we know that people vary in their scores from one repetition of the test to another by as much as 10 points and the difference between the scores of two persons is 30 points, we probably should be willing to conclude that one person is indeed superior to the other and the difference would hold even if we tested them on another occasion . We should know that if we administered the test a second time to the same two people, the individual who was superior on the first occasion undoubtedly would be superior on the second . The one who was superior on the first application of the test might earn a score as much as 10 points lower on the second test and the one who was inferior on the first test might improve his score by as much as 10 points, but there would still be at least a 10-point difference between the two individuals in their scores .

However, we should not be so willing to say that the one person is superior to the other if we found that people vary as much as 20 points from one application of the test to another . In this case, if we tested both persons twice, the individual who earned the higher score on the first application of the test might well earn the lower score on the second application .

D2

Reliability of measurement is an important consideration in terms of the precision with which individuals can be assigned to groups or classes .

Suppose that pupils in a school are to be placed in reading sections on the basis of the scores they earn on a reading achievement test , with those earning scores of 60 and above being assigned to the accelerated section, those with

scores of 50 to 59 to the average section, and those with scores of 49 and below to the retarded section. Now further suppose that variation in scores of as much as six points occurs when an individual takes the test a number of times. A pupil who earns a score of 55 on the test will be assigned to the average section. However, if he had taken the test on another occasion he might have earned a score as high as 61 and have been assigned to the accelerated section, or he might have earned a score as low as 49 and have been assigned to the retarded section. Because of the degree of unsystematic variation in individual's scores the degree of reliability of measurement of this test is insufficient to assign pupils to sections with very much certainty. On the other hand if the variation among scores in subsequent administrations of the test is only one point, then a very large proportion of pupils can be assigned with a high degree of certainty.

D3

The accuracy of prediction from one variable to another is limited by the degree of reliability with which these variables are measured. Scores on tests often are used to predict other types of behavior. For example, scores on intelligence tests commonly are used to predict success in academic work. If the particular intelligence test used in making predictions of this kind happens to be highly unreliable, then from the score earned by an individual on one occasion a high degree of scholastic success might be anticipated for him, but from the score he earned on another occasion just the opposite conclusion might be reached. It would, therefore, be difficult under such circumstances to make

predictions with any satisfactory degree of certainty .

D4

The confidence we place on the differences among the scores earned by an individual on different tests is a function of the degree of reliability of those tests . In certain circumstances it is necessary to know on which of two traits an individual is superior . For example , as an aid to counseling a student we may wish to know whether he is superior in mechanical or in clerical aptitude . Suppose when we apply tests of mechanical and clerical aptitude to the same individuals , scores in each test vary as much as 20 points from one application to another . If we administer both tests to a student on a single occasion and find that on the mechanical aptitude test his score is 70 and on the clerical aptitude test his score is 85 , we can not say with much certainty that his mechanical aptitude is superior to his clerical aptitude . However if the variation in scores on repeated applications of the tests were only five points , we should be much more willing to draw this conclusion .

D5

Sub-D1

As stated previously since no testing instrument is perfectly reliable , errors and variations in measurement will occur . There are two major types of variation in scores earned by an individual over repeated testing . Systematic variation is characterized by a systematic change in score , while unsystematic variation is characterized by random and unsystematic fluctuations in scores . When scores exhibit unsystematic variation the test is not measuring accurately and has low reliability . It means that we cannot depend too much upon any single score earned by an individual since on another application

of the same test he might earn quite a different score. We must examine both types of variation so that we can differentiate unsystematic from systematic variation in order to further develop the concept of reliability. D6

Systematic variations in scores are characterized by an orderly progression or pattern, with the scores obtained by an individual changing from one occasion to the next in some trend. Systematic changes appear as a regular increase or decrease in scores or they may appear to follow some cycle. Suppose we examine the different scores obtained by an individual on successive applications of the same test. A trend might appear. Thus if we measure the height of an individual at different hours of the day we are likely to find that from the morning to evening the values become smaller and smaller. We might attribute this phenomenon to a gradual sagging of the backbone. Similarly if we administer the same arithmetic test over and over to the same individual, his scores may gradually increase. This would suggest that the series of testing situations operate as practice periods and the individual is gradually improving his skill in solving arithmetic problems.

Unsystematic variation, on the other hand, is characterized by a complete lack of order. The scores of an individual fluctuate from one occasion to the next in a completely haphazard manner. For example, if we have an individual react as quickly as possible in a specified manner to each stimulus in a series of stimuli, we shall find that some of his responses are more rapid than others. When we compare the times taken to respond to stimuli that occur

early in the series with those that occur later, we may find that on the average they are the same. We might attribute this variation in reaction-time scores to unsystematic moment-to-moment changes in the environmental conditions, in the smoothness of the operation of the reaction-time apparatus, in the individual's motivation, and in his attention. D7

Various factors can influence a particular kind of variation in test scores. These factors can be classified as systematic or unsystematic according to whether their effects on test scores are systematic or unsystematic.

A systematic factor is one which produces systematic changes in scores. When systematic factors are at work, scores show a regular arrangement, an order. Learning, training, and growth produce regular and progressive increases in scores. Fatigue, forgetting and old age result in regular and progressive decreases in scores. Mood and living habits may produce regular cyclical changes in scores.

An unsystematic factor is one which produces unsystematic changes in scores. Scores fluctuate in a random fashion and do not manifest any consistent pattern. Moment-to-moment variations in attention result in random fluctuations in reaction time. An inconsistent and balky pen sometimes permits the student taking an exam to write easily and on other occasions slows the speed of writing. The marks given to an elementary school pupil as he progresses through the various grades are sometimes higher and sometimes lower, depending upon whether the teacher to whom he happens to be assigned tends to be lenient

or strict in the evaluation of pupils' performance .

D8

The factors which affect scores seem to be almost infinite in number and variety . An individual's performance is a function of the numerous qualities with which he was endowed at birth , elaborated upon by the process of maturation and by his numerous experiences , together with the many environmental influences operating upon him at any given moment . The inferences we draw in attempting to explain variation in scores are a function of the knowledge we have about these factors . In some instances our inferences have quite substantial foundation because our knowledge is direct and extensive . In other instances our knowledge may be indirect and not complete so that we are less sure of our inferences . Finally we may have such limited knowledge about conditions that our inferences are little more than guesses .

In the following situation we can be fairly sure of the factors which are operating . Suppose we give an individual a test of knowledge of French vocabulary and find his score is zero . We then have him take an elementary course in French and retest him . Now his score is higher . He continues to take more and more courses in French and after he completes each course we again administer the test . Undoubtedly we shall find a continuous increase in his scores , and with a high degree of certainty attribute his increase in scores to the training to which he has been deliberately subjected .

At the other extreme , under some conditions inferences about operative factors may be guesses . Suppose we have before us determinations

of the intelligence of a child from several different testings and note that they were substantially lower when the tests were administered during the summer months. We have no knowledge whatsoever about the state the child was in when he was tested nor of the conditions prior to or coincident with the different administrations of the test. There are a variety of inferences we might make to explain the variation in scores. One which might appear reasonable to us is that this child's performance on intelligence tests is influenced by the degree of intellectual stimulation he receives, so that during the summer months when he is away from school his scores are lower. This accounts for the changes but is only a guess.

Having examined the kinds of variation that occur in scores and the type of factors that cause them, we are now in a position to define reliability more precisely. We shall do so in terms of the extent of unsystematic score variation and the concept of parallel tests.

Let us refer to reliability as the extent of unsystematic variation in the quantitative description of the amount of some trait an individual possesses or manifests when that trait is measured a number of times. This definition follows from the fact that the problem of reliability of measurement arises out of the unsystematic variation in scores earned by an individual when we obtain a number of measurements indicative of the degree to which he possesses or manifests some particular trait or quality. Therefore, reliability of measurement pertains to the precision with which some trait is measured by

means of specified operations .

Basic to any formal mathematical statement of reliability is the concept of parallel tests . In essence , reliability can be defined as the extent of unsystematic variation of an individual's scores on a series of parallel tests . Parallel tests refer to a number of operations or tests all of which follow from a particular definition of a trait , and therefore measure the same trait to the same degree . Certain statistical criteria must be met in order that a given trait is measured exactly the same . Theoretically to ascertain the extent of unsystematic variation parallel tests are needed , i.e. , a series of scores on the same trait . It is not necessary to always use the same device or test , nor do we have to deny usage of the same measuring device or test . All that is needed is tests or operations which evoke the same psychological processes . D9

It is through the use of parallel tests that we are able to ascertain the extent to which we are measuring a trait reliably . Suppose we have a series of parallel tests , k in number and we have scores on all these k tests for one or more individuals . If we were measuring with perfect reliability then any given individual would obtain precisely the same score on all the k parallel tests . There would be no variation at all in his scores over the k tests . On the other hand , if we were measuring with less than perfect reliability then his scores would be different on the different parallel tests , the variation among his scores being completely unsystematic . The less the unsystematic variation the greater the reliability of measurement and the greater the unsystematic variation the

less the reliability of measurement .

We have seen that reliability of measurement refers to the extent of unsystematic variation in an individual's scores over parallel tests . The next task is to set up an index which gives a quantitative description of the extent of such variation . Such an index will be useful for comparing different tests so we can ascertain which gives us the most precise or stable scores , and will permit us to ascertain whether the reliability with which a test measures is sufficient for our purposes .

Theoretically the reliability coefficient is a quantitative index of the extent to which scores on any one parallel test can predict scores on any other . When the unsystematic variation in an individual's scores over parallel tests is great , this means that the prediction of scores on one parallel test from scores on another is poor . On the other hand , if there is no unsystematic variation at all among an individual's scores , then it means that we could predict perfectly an individual's score on one parallel test from his score on another .

Casting reliability in terms of the coefficient of correlation between parallel tests provides a quantitative way of describing the precision of measurement . Essentially , a correlation coefficient expresses the degree of correspondence or relationship between sets of scores . A correlation coefficient can be computed between sets of scores from any combination of tests , one may be an achievement test and another a personality measure .

However, when defining the reliability coefficient we restrict the classification of tests that are correlated to parallel tests as defined previously. We then define the reliability coefficient as the correlation coefficient between parallel tests. When the correlation coefficient is low it means that an individual's scores over k parallel tests show a great deal of unsystematic variation, and when it is high it means that an individual's scores on k parallel tests are very nearly the same. Let us consider the concept of correlation further.

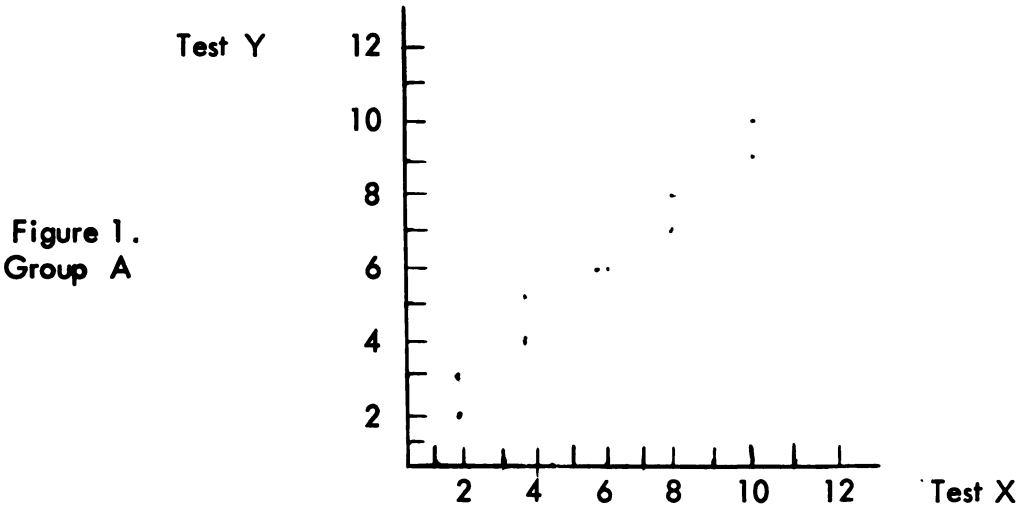
D10

Sub D-2

Suppose we have a set of scores from a group of individuals (A) on a couple of tests, represented by the symbols X and Y , which are as follows:

<u>Group A Persons</u>	<u>Scores on Test X</u>	<u>Scores on Test Y</u>
1A	2	3
2A	2	2
3A	4	4
4A	4	5
5A	6	6
6A	6	6
7A	8	7
8A	8	8
9A	10	9
10A	10	10

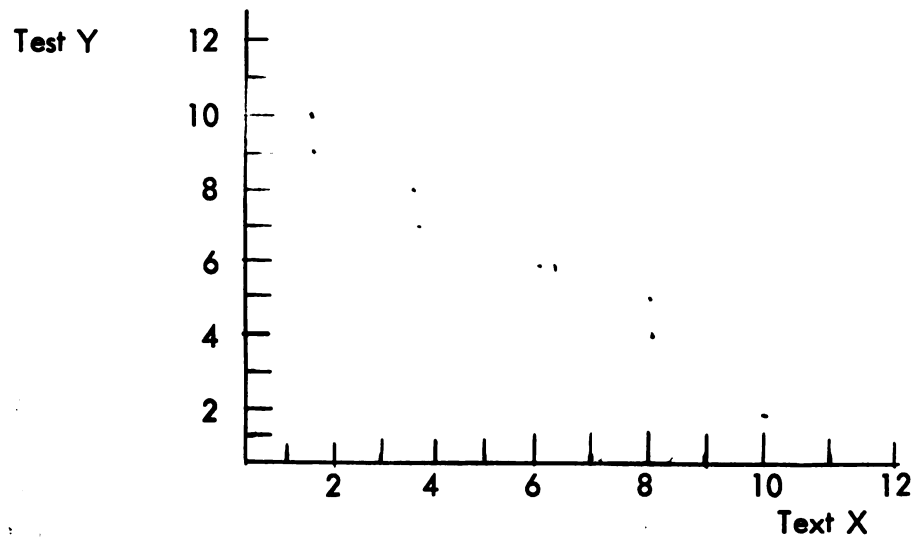
If we plot the scores of each person on these two tests on a graph, where each point represents the two scores of an individual, one on Test X and the other on Test Y , we then have the scatter diagram in Figure 1.



If we use the same tests X and Y on another group of individuals (B), we might obtain the following set of scores and scatter diagram (Figure 2).

<u>Group B Persons</u>	<u>Scores on Test X</u>	<u>Scores on Test Y</u>
1B	2	9
2B	2	10
3B	4	8
4B	4	7
5B	6	6
6B	6	6
7B	8	4
8B	8	5
9B	10	2
10B	10	3

Figure 2,
Group B

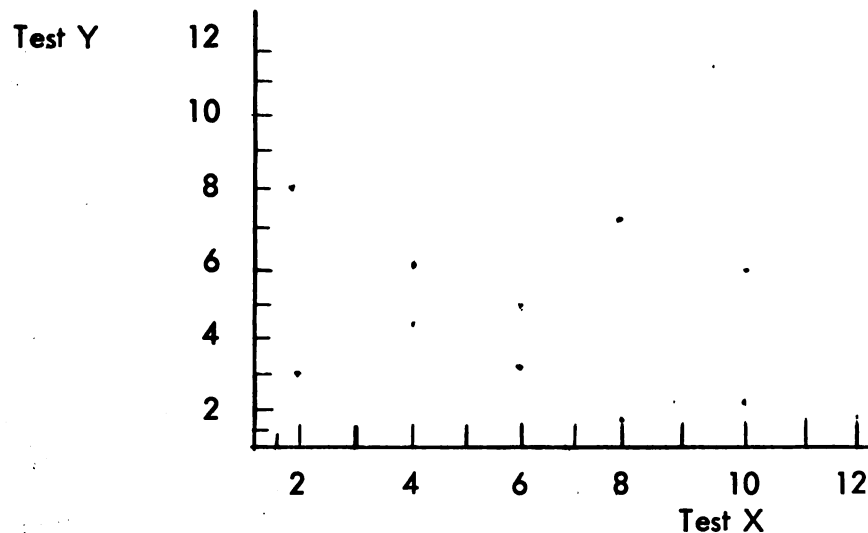


Following the same procedure with another group of individuals

(C) we might obtain another set of scores and scatter diagram (Figure 3).

Group C Persons	Scores on Test X	Scores on Test Y
1C	2	3
2C	2	8
3C	4	6
4C	4	4
5C	6	3
6C	6	5
7C	8	7
8C	8	1
9C	10	2
10C	10	6

Figure 3.
Group C



With group A the order of individuals on one test is quite similar to their order on the other test. That is, if an individual scores high (low) on one test he scores high (low) on the other. With group B the order of the individuals on one test is practically the reverse of the order on the other test; if an individual scores high (low) on one test he scores low (high) on the other. However, with group C the order of individuals on one test is not at all similar to the order on the other test; if an individual from group C scores high (low) on one test, he could score either high or low (low or high) on the other test (refer back to the appropriate scatter diagrams for clarification if needed).

We can say then that there is a high relationship or correspondence between scores on the two tests, X and Y, for groups A and B, but a low relationship between them for group C. That is, for groups A and B if the

score of an individual on one test is known his score on the second test can be predicted with a high degree of accuracy. But for group C such a prediction would be quite subject to error. The correlation coefficient is the quantitative measure which reflects this degree of relationship between sets of scores. It is symbolized by r_{xy} where x and y refer to the two correlated tests. The correlation coefficient itself can range numerically from $+1.00$ to -1.00 (both termed high correlation coefficients) where $+1$ and -1 represent perfect relationships between two tests similar to that illustrated in the scatter diagram of groups A and B respectively. On the other hand a correlation coefficient of 0.00 reflects no relationship at all between two tests, such as that illustrated by the scores and scatter diagram of group C.

It will be recalled that the correlation coefficient between parallel tests is termed the reliability coefficient. As such a high correlation coefficient means a high reliability coefficient and a low correlation coefficient means a low reliability coefficient. Now let us assume that for groups A and C the tests X and Y represent parallel tests measuring the same trait. As stated before there is a high correlation between X and Y for group A. In other words, there is little unsystematic variation in scores over the parallel tests for group A resulting in a high correlation coefficient and, therefore, a high reliability coefficient. Given a specific score on test X for group A the variation of scores on Test Y is quite small (little unsystematic variation), therefore, the correlation between test X and Y is high for group A. For example, if an individual

in group A receives a score of 8 on test X, then he is apt to receive a score between 7 and 8 on test Y (see page 228 for scatter diagram).

However, the situation is different for group C on test X and Y. Here there is a low correlation between scores; there is much unsystematic variation over the parallel tests, lowering the correlation coefficient and hence the reliability coefficient for this group. Given a specific score on test X, the variation of scores on test Y is quite large (much unsystematic variation), therefore, the correlation between tests X and Y is low for Group C. For example, if an individual in group C receives a score of 3 on test X he may receive a score anywhere between 1 and 7 on test Y (see page 230 for scatter diagram).

D11

/Formal Break/

In order to develop our notions of reliability so that we can consider practical ways for measuring its extent, we shall have to examine in more detail the nature and effects of unsystematic factors. Previously we have distinguished between consistent trends in scores that are attributable to unsystematic factors. We have defined reliability theoretically in terms of the extent of unsystematic variation in individual's scores over repeated testing. Now we can separate the class of unsystematic factors into two types, varying unsystematic factors and constant unsystematic factors.

D12

Sub - D3

Varying unsystematic factors refer to those whose effects are different for the same individual on different occasions and are also different

for different individuals on the same occasion. Constant unsystematic factors refer to those whose effects are different for the same individual on different occasions but are the same for all individuals on the same occasion. Thus the difference between these two types of unsystematic factors lies in their effects in a single testing occasion. Let us examine varying unsystematic factors first.

The effects of varying unsystematic factors are different for the same individual on different occasions. Hence the score of an individual over a number of occasions is sometimes higher and sometimes lower as a result of varying unsystematic factors. In addition, the effects of these factors are different for different individuals on the same occasion, tending to increase the scores of some individuals on that occasion and to lower those of others. Some of these influences are in the testing situation itself and others are ascribed to the individual.

Let us first examine the different sources of varying unsystematic variation that are in the testing situation itself. For example, some persons may be fortunate enough to sit in comfortable seats while they are taking a test, whereas others may find themselves in uncomfortable seats. Those near the window work under conditions of good illumination whereas others may find themselves in the far corners, operating under the handicap of poor lighting. Because of their nearness to or distance from the test administrator, some hear the instructions clearly and others do not. These situations illustrate that

varying unsystematic factors have different effects on different individuals on the same occasion. When the measuring instrument is a rating procedure, differences among raters produce variations in scores. The individual who happens to be rated by a lenient rater is likely to receive a higher rating than one who is rated by a strict rater. Assuming the rating is done at different times, this situation illustrates that varying unsystematic factors have different effects on the same individual over different occasions.

Because the individual himself changes in unsystematic ways, he too is a source of random variation of this type. His motivation, fatigue, nervousness, interest, and distractibility may be to one degree on one occasion and to another degree on another occasion, thus having different effects on the same individual over many occasions. Different individuals taking a test at a given time also vary among themselves in these same respects. The motivation of some happens to be high at the time of testing, whereas that of others happens to be low; and some people happen to be rested, whereas others happen to be tired, illustrating varying unsystematic factors having different effects on different individuals on the same occasion. D13

The other type of unsystematic factor, constant unsystematic, is unsystematic in its general influence over a number of testing occasions but yet operates in the same fashion for all individuals at a given time. When constant unsystematic factors are operating the scores of all individuals on one occasion may be higher or lower than their scores on another occasion,

with the scores varying in a non-orderly, random fashion. For example, we may have a speed test with a 10 minute time interval. Sometimes through erroneous reading of the timing device the test administrator may shade the 10-minute interval by several seconds, while on other occasions he may unknowingly be several seconds too generous in his timing. On any given occasion the time, though in error, is the same for all individuals being tested. Sometimes when a test is administered the lighting may be poor throughout the entire testing room and on another occasion it may be excellent. While from one testing session to the next there are variations in quality of illumination, on any one occasion it is the same for all individuals. When individuals are given a test in the morning they may all be fresh and rested and consequently earn higher scores on it than when they take the test in the evening and are all tired. In these situations on any one occasion constant unsystematic factors effect all individuals in the same manner, e.g. all had time cut short, all had poor lighting. Yet across many occasions the effects of constant unsystematic factors are different for the same individual, varying in an unsystematic fashion. For example the lighting conditions over many occasions may be as follows: poor light, really bad light, excellent, almost dark, mediocre, horrible light, etc. - with no predictable order as to the exact lighting condition on any given day.

D14

As stated before the primary difference between constant and varying unsystematic factors lies in their effects in a single testing occasion;

constant unsystematic factor effects cannot be detected on one occasion because they have the same effect on all individuals, on that occasion, while varying unsystematic factors can be detected since they have different effects on different individuals thus producing random variation in the scores. However, over different occasions both types of unsystematic factors have different effects on an individual's scores. This latter condition, in fact, is exactly what they have in common and why they are both labelled unsystematic. This distinction between constant and varying unsystematic factors needs to be specified further.

It is clear that both types of factors produce unsystematic, non-orderly variations in an individual's scores on different occasions. Therefore, if we administer a test to a single individual on many occasions, we cannot distinguish between the two types of factors on the basis of the scores alone. Suppose we administer a test to only one person at a time but to each person we administer the test a number of times. If we have n people and we administer the test k times to each person, we then have nk occasions on which the test has been administered. On each of these nk occasions the effects of varying unsystematic factors are different and also the effects of constant unsystematic factors are different. Hence we could not distinguish their effects and we could not ascertain which type of factor is determining variation among scores or whether both are at work. If the various testings of the subjects are randomly distributed among the k occasions the constant factors operate in exactly the same manner as the varying factors, because the constant unsystematic factors

do not have the same constant effects upon all individuals, i.e., each individual being tested on a different occasion rather than all tested on the same occasion. Therefore, all individuals have equal likelihood of being tested under favorable and unfavorable conditions of constant as well as varying factors.

We might conclude that constant factors should be classified with systematic rather than with unsystematic factors; since they overlap in function, operating similarly on the same occasion, i.e., both having the same effect on all individuals on that occasion. However, they are different. The reason for this separate classification is that constant unsystematic factors cause the scores of an individual to vary in a random and unpredictable fashion from occasion to occasion, whereas systematic factors produce systematic and predictable changes over occasions. Since from one occasion to another constant unsystematic factors operate in a random fashion for a given individual they are classed as unsystematic. However, systematic factors have the same effects on an individual on different occasions. To illustrate this, if a systematic factor has a facilitating effect on one occasion it will also have a facilitating effect on the following occasions. For example, if tests are always given at the same hour in the morning and we assume that students are fresh at this time; then time of day is a systematic factor affecting everyone the same on one occasion and across occasions as well.

D 15

Sub - D4

We have defined reliability as the extent of unsystematic variation

of an individual's scores over a series of parallel tests. The quantitative index of this amount of variation is the reliability coefficient, i.e., the correlation coefficient over parallel tests. In a practical situation we rarely have parallel tests available but we usually do have several means of administering non-parallel tests by which we can estimate the reliability coefficient. Any estimation procedure will give us only an approximation, not an exact determination, of the reliability coefficient. There are three basic methods that are used for estimating the reliability coefficient of tests. They are (a) test-retest: estimation from the correlation coefficient between scores on repetitions of the same test, (b) parallel forms: estimation from the correlation coefficient between scores on parallel forms of a test and (c) internal consistency: estimation from correlation coefficient among comparable parts of the test. In the discussion below only group testing, not individual, is considered.

D16

The first method for estimating the reliability coefficient is called the test-retest method. A certain test is administered two or more times to the same group of individuals, and the intercorrelations among the scores on the various administrations are taken as the reliability coefficient. With tests of aptitude, personality, and achievement the test ordinarily is administered only once so that only one estimate of the reliability coefficient is obtained. If the test is administered several times, the usual practice is to take the average of the intercorrelations among the scores on the various occasions as the estimate

of the reliability coefficient .

D17

There are two main advantages with the test-retest method . Nothing in addition to the test itself is required . The particular sample of items or stimulus situation is held constant , thereby testing the individuals with precisely the same instrument .

The most serious disadvantages with the test-retest method lie in the variety of carry-over effects from one testing occasion to another . Sometimes there are practice effects so that on subsequent occasions scores increase in a systematic fashion . The individual may learn the specific content of the test or develop improved approaches or attitudes toward the material so that his scores increase . In some instances these practice effects are different for different individuals . Of two people who obtain precisely the same score on the first occasion , one may discover certain general principles that help answer the questions in the test or may even rehash or rehearse the material during the interval between the first and second testing . Therefore , on the second testing occasion the scores of one individual may be improved and that of the other may remain the same . If the correlation between the scores on the two occasions is low , we do not know whether the test is unreliable or whether differential systematic factors have been at work . On the other hand , if the coefficient between scores on the two occasions is high , then it would seem that factors having differential effects are not very important and the correlation we obtain might be considered to be something like a lower limit of the reliability

coefficient. This would be true, of course, only if we could rule out on the retest the effects of remembering the response made on the first test.

In other instances there might be a specific carry-over effect in terms of remembering on one testing occasion the response given on an earlier one and merely repeating these responses. In an attitude test on the first testing occasion a person answers "indifferent" to the question "Do you approve of labor unions?" and remembering this on a second occasion, he again responds in the same fashion. Having assigned his subordinate Joe Smith the rating of "superior" in January, a factory foreman does so again in June when he is called upon to rate him in order to demonstrate that he is consistent in his appraisals. These specific carry-overs from one occasion to another may not be deliberate on the part of the individual; indeed, he may be completely unaware of them. Their presence in the test-retest method may give an overestimate of reliability. They introduce a false consistency in scores.

One troublesome problem with the test-retest method has to do with the time interval between testing occasions. We expect lower and lower estimates of reliability as the time interval between the testing occasions increases, because the longer the time interval between the two testing occasions the greater the likelihood that the individual will change. Yet in order to minimize the effects of memory, it is desirable to maximize the interval between testing occasions. Therefore, the correlation between scores on two occasions reflects the ability of individuals to remember, as well as the reliability of measurement.

The second method of estimating reliability is that of parallel forms. Parallel forms of a test should not be confused with parallel tests. Parallel forms of a test are tests similar in content and nature designed to measure the same traits. Parallel tests are not necessarily similar in content and nature. As stated before, parallel tests measure the same trait and must meet certain statistical criteria, which we have not specified here. If a series of parallel forms of a test meet these criteria they are also parallel tests. But if they do not, they are only parallel forms. To illustrate the concept of parallel forms of tests as tests which are similar in content or nature, consider these examples. Two objective tests might have the same kind and number of items. An item in one parallel form of an arithmetic test might be " $27 + 83 = \underline{\quad}$ ", and an item in another form might be " $48 + 72 = \underline{\quad}$ ". An item in one form of an inventory designed to measure emotional stability might be "Do you sleep well at night?" and an item in another form might be "Do you have bad dreams at night?"

Sub-D5

Having available two or more parallel forms of a test, we take as an estimate of the reliability the intercorrelations among the scores on the parallel forms. If there are more than two forms available the common practice is to take the average of the intercorrelations as the estimate of the reliability coefficient. The intercorrelations among the tests reflect not only the degree of reliability of measurement but also the extent to which the tests measure different traits, since the various forms of a test do not contain precisely the

same material. Hence we might say that the method of determining reliability from the intercorrelations among parallel forms of a test gives estimates that are too low.

The carry-over effects from one test to another are minimized because the content of parallel forms is not precisely the same. In many instances there will be no specific carry-over effects at all, because there is no opportunity to memorize specific responses made to an earlier form. However, there is still the possibility of general carry-over in terms of modes of response, attitudes toward the material and the like. Ordinarily when the method of parallel forms is used to estimate reliability of measurement, the various forms are administered on different occasions, termed parallel forms-delayed; although sometimes they are administered on the same occasion, termed parallel forms-immediate.

D18

The last method of estimating reliability, internal consistency, involves only a single administration of a test. Under such circumstances we can obtain an estimate of the reliability of measurement if we consider the test not as a single test but rather as the sum total of a number of parallel forms of a test. Suppose we have an objective test comprised of 100 items all of which pertain to the same trait. Instead of saying that we have one test of 100 items we might say that we have two tests each of 50 items or four tests each of 25 items or 100 tests each consisting of one item. Having two or more parallel forms available, we can now proceed to estimate reliability coefficients

by the method of correlation between scores on parallel forms. Note that we do not have the reliability of a test of 100 items but rather the reliability of a shorter test. If we do not feel that the shorter tests adequately sample the trait we wish to measure, we can find the reliability coefficient of the total test by various statistical methods.

Usually the test is divided into two parts. A problem arises about splitting the test. With a 100 item test we could take the first 50 items as one half and the last 50 as the other half, or we could take the odd-numbered items as one half and the even-numbered ones as the other half. This last procedure, the odd-even method, is the one generally used since it controls for any systematic factors operating during the testing period that change the performance from early in the testing session to later periods; an example of such a factor is fatigue. In order to maximize the probability that the two halves measure the same trait sometimes the division is made on the basis of an analysis of the content of the items, making sure that both halves contain items of the same sort.

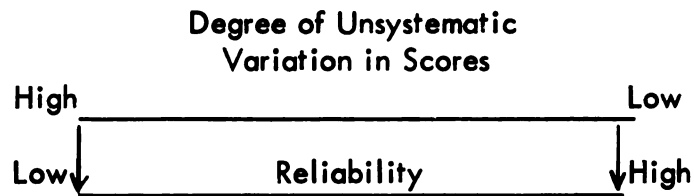
The prime advantage of determining reliability coefficients by this method is its simplicity. A test need be given only once to a group of individuals; repetition of the test or parallel forms are not required. The method is not applicable to certain types of tests which are an integrated whole and cannot be divided into separate and equivalent parts, as is the case with speed tests.

We have developed the concept of reliability both theoretically and practically. We have seen that reliability plays an important role in the practical application of test results. Yet it serves only as a necessary not a sufficient condition for quality in a test. That is, we could be measuring something with high reliability but which is trivial. On the other hand, if a satisfactory reliability is not achieved nothing has been measured very precisely.

Review Diagram

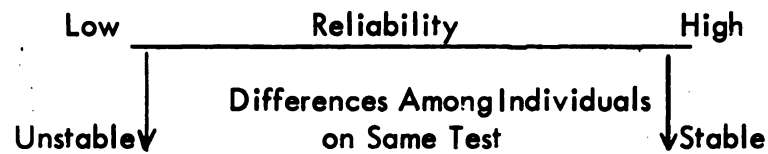
APPENDIX F
DIAGRAMS AND
CORRESPONDING VERBAL STATEMENTS

D1



The degree of unsystematic variation in scores and the degree of reliability may range from low to high; a high degree of unsystematic variation yielding low reliability and low degree of unsystematic variation yielding high reliability.

D2



Reliability may range from low to high and differences among individuals on the same test range from unstable to stable; low reliability yielding unstable differences while high reliability yields stable differences.

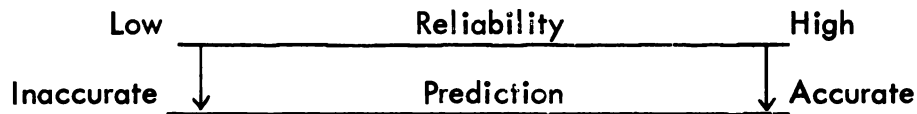
D3



Reliability may range from low to high and assignment of

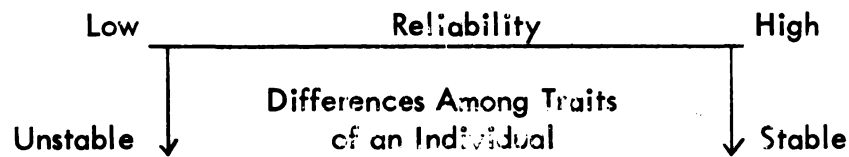
individuals to groups range from uncertain to certain; low reliability yielding uncertain assignment while high reliability yields certain assignment.

D4



Reliability may range from low to high and prediction range from inaccurate to accurate; low reliability yielding inaccurate prediction while high reliability yields accurate prediction.

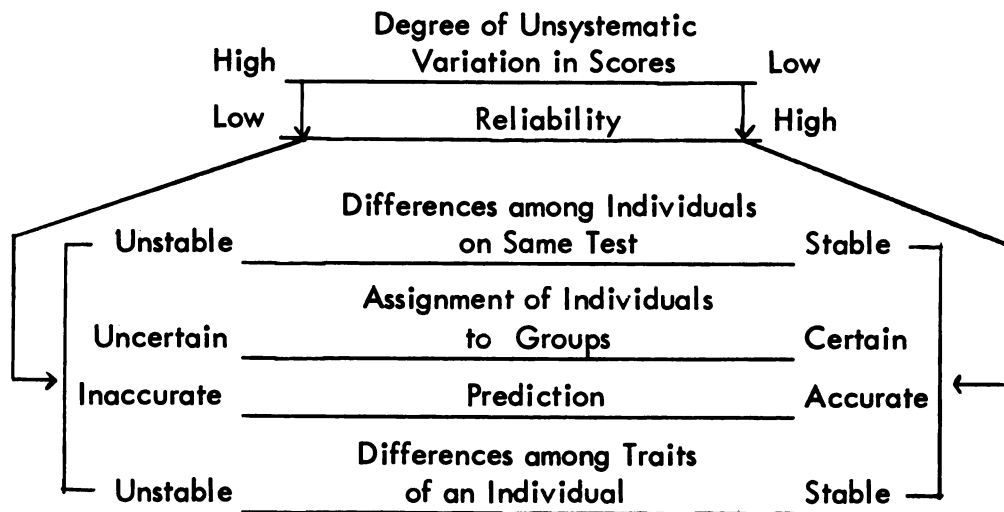
D5



Reliability may range from low to high and differences among traits of an individual range from unstable to stable; low reliability yielding unstable differences while high reliability yields stable differences.

Sub-D1

In summary then, we have the following:



The degree of unsystematic variation in scores and reliability can range from low to high. A high degree of unsystematic variation yields low reliability, unstable differences among individuals on the same test, uncertain assignment of individuals to groups, inaccurate prediction and unstable differences among traits of an individual. A low degree of unsystematic variation yields high reliability, stable differences among individuals, certain assignment, accurate prediction, and stable differences among traits.

D6

Types of Variation in Scores



There are two types of variation in scores, systematic and unsystematic.

			Reliability of Measurement	
Measure Same Trait	X	Parallel Tests	Extent of Unsystematic Variation in Individual's Scores	X
Meet Statistical Criteria	X		Parallel Tests	X

D9 (Continued)

Reliability of measurement is the extent of unsystematic variation in an individual's scores over parallel tests. Parallel tests measure the same trait and meet statistical criteria.

D10

		Reliability Coefficient	Correlation Coefficient
Quantitative Index		X	X
Degree of Unsystematic Variation in Individual's Scores		X	X
Tests	Parallel	X	X
	Not-parallel		X

The reliability coefficient is a quantitative index of the degree of unsystematic variation in an individual's scores over parallel tests. The correlation coefficient is a quantitative index of the degree of unsystematic variation in an individual's scores over parallel and not-parallel tests.

Sub-D2

In summary, we can then describe reliability, reliability coefficient, and correlation coefficient in terms of the following characteristics:

Sub-D2 (continued)

		Reliability	Reliability Coefficient	Correlation Coefficient
Quantitative Index			X	X
Degree of Unsystematic Variation in Individual's Scores		X	X	X
Tests	Parallel	X	X	X
	Not-parallel			X

Reliability is the extent of unsystematic variation in an individual's scores over parallel tests. The reliability coefficient is a quantitative index of this degree of unsystematic variation; the correlation coefficient is a quantitative index of the degree of unsystematic variation in an individual's scores over parallel and not-parallel tests.

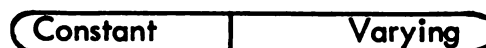
D11



The correlation coefficient on parallel tests is the same as the reliability coefficient, both ranging similarly from low to high.

D12

Types of Unsystematic Factors



There are two types of unsystematic factors, constant and varying.

Sub-D3

At this point we can now briefly re-examine the types of factors and their respective outcomes (with the diagram given below).

		Systematic Factors	Unsystematic Factors	
		Systematic Variation	Constant	Varying
			Unsystematic Variation Constant	Unsystematic Variation Varying
Type of Score Arrangement	Orderly Pattern	X		
	Complete Lack of Order		X	X

Systematic factors cause systematic variation in scores, characterized by an orderly pattern of scores; and unsystematic factors cause unsystematic variation characterized by complete lack of order. Constant and varying are the two types of unsystematic factors yielding constant and varying unsystematic variation respectively.

Now let us examine the two types of unsystematic factors further.

D13

		Effects		Individual(s)		Occasion	
		Same	Different	Same	Different	Same	Different
Varying Unsystematic Factors			X		X	X	
			X	X			X

(Read across each row.)

D13 (continued)

Varying unsystematic factors have different effects on different individuals on the same occasion and have different effects on the same individual on different occasions.

D14

	Effects		Individual(s)		Occasion	
	Same	Different	Same	All	Same	Different
Constant Unsystematic Factors		X	X			X
	X			X	X	

Constant unsystematic factors have different effects on the same individual on different occasions and have the same effect on all individuals on the same occasion.

D15

	Effects		Individual(s)		Occasion	
	Same	Different	Same	All	Same	Different
Systematic Factors	X			X	X	
	X		X			X

Systematic factors have the same effect on all individuals on the same occasion and have the same effect on the same individual on different occasions.

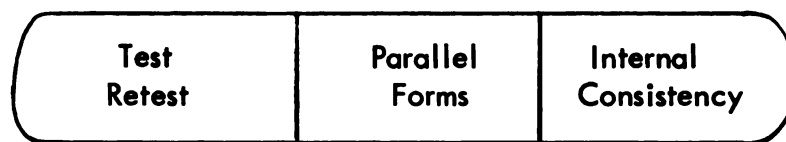
Sub-D4

		Systematic Factors		Unsystematic Factors		
				Constant	Varying	
Effect	Same	X	X	X		
	Different				X	X
Individuals	Same	X			X	
	Different					X
	All		X	X		
Occasion	Same		X	X		X
	Different	X			X	

Systematic factors have the same effect on all individuals on the same occasion and the same effect on the same individual over different occasions. Constant unsystematic factors also have the same effect on all individuals on the same occasion but have different effects on the same individual on different occasions. Finally, varying unsystematic factors have different effects on different individuals on the same occasion and different effects on the same individual on different occasions.

D16

Methods of Estimating the Reliability Coefficient



There are three main ways of estimating the reliability coefficient; test-retest, parallel forms, and internal consistency.

D17

	Type of Test Administered		Time of Testing		Number of Times Test Admin.	
	Identical	Similar	Same Occasion	Different Occasion	Once	More than Once
Test Retest	X			X		X

The test-retest method involves administering the identical test on different occasions.

Sub-D5

	Parallel Tests	Parallel Forms of a Test
Always Similar in Content		X
Measure Same Trait	X	X
Meet Statistical Criteria	X	

Parallel tests measure the same trait and meet statistical criteria. Parallel forms of a test are always similar in content and measure the same trait.

D18

		Type of Test Administered		Time of Testing		Number of Times Test Admin.	
		Identical	Similar	Same Occasion	Different Occasion	Once	More than once
Parallel Forms	Del.		X		X		X
	Imm.		X	X			X

The parallel forms-delayed method involves administering similar tests on different occasions. The parallel forms-immediate method involves administering similar tests on the same occasion.

D19

		Type of Test Administered		Time of Testing		Number of Times Test Admin.	
		Identical	Similar	Same Occasion	Different Occasion	Once	More than once
Internal Consistency		X		X		X	

The internal consistency method involves administering the identical test once.

Sub-D6

All of these methods of estimating reliability coefficients can be characterized as follows:

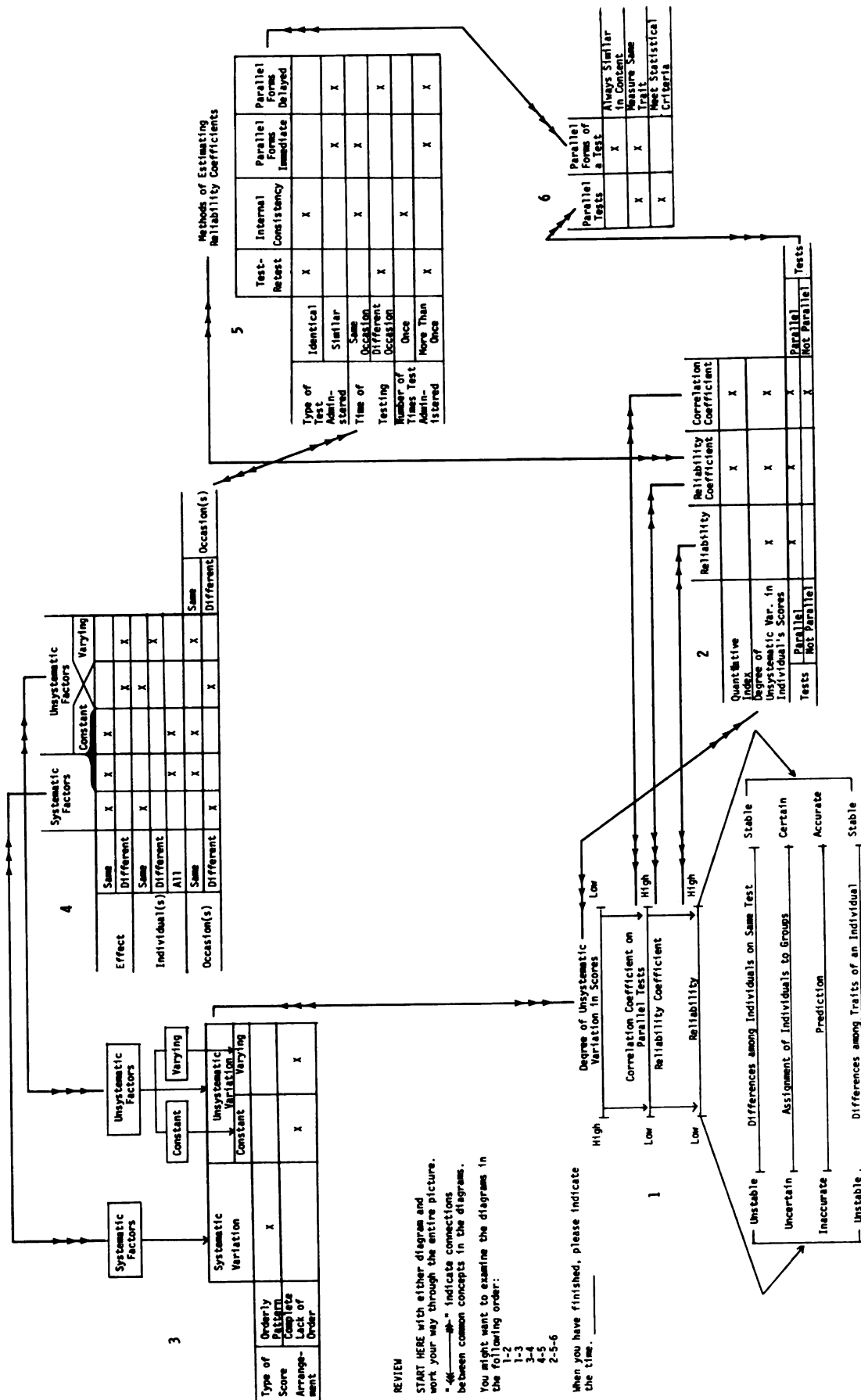
Sub-D6 - continued

		Test Retest	Internal Consistency	Parallel Forms Immed .	Parallel Forms Delayed
Type of Test Administered	Identical	X	X		
	Similar			X	X
Time of Testing	Same Occasion		X	X	
	Different Occasion	X			X
Number of Times Test Administered	Once		X		
	More than Once	X		X	X

Test-retest involves administering the same test on different occasions; internal consistency method involves giving a test once; parallel forms-delayed involves administering similar tests on different occasions; parallel forms-immediate involves administering similar tests on the same occasion .

Instructions preceding the review diagram

On the next page is a large diagram which reviews and integrates the diagrams presented in the test into six sub-diagrams . Interconnections between concepts in these sub-diagrams are indicated . The sub-diagrams are numbered and a suggested order of progressing through the entire diagram is given . While interpreting the chart diagrams , read down the columns rather than across the rows .



(Verbal Review)

The degree of unsystematic variation in scores, correlation coefficient on parallel tests, reliability coefficient and reliability can range from low to high. The correlation coefficient on parallel tests is the reliability coefficient. A high degree of unsystematic variation yields low correlation coefficients, low reliability coefficients and low reliability, while a low degree of unsystematic variation yields high correlation coefficients, high reliability coefficients and high reliability. Low reliability yields unstable differences among individuals on the same test, uncertain assignment of individuals to groups, inaccurate prediction, and unstable differences among traits of an individual. High reliability yields stable differences among individuals, certain assignment, accurate prediction and stable differences among traits.

Systematic factors cause systematic variation in scores, characterized by an orderly pattern of scores; and unsystematic factors cause unsystematic variation characterized by complete lack of order in scores. Constant and varying are the two types of unsystematic factors yielding constant and varying unsystematic variation respectively.

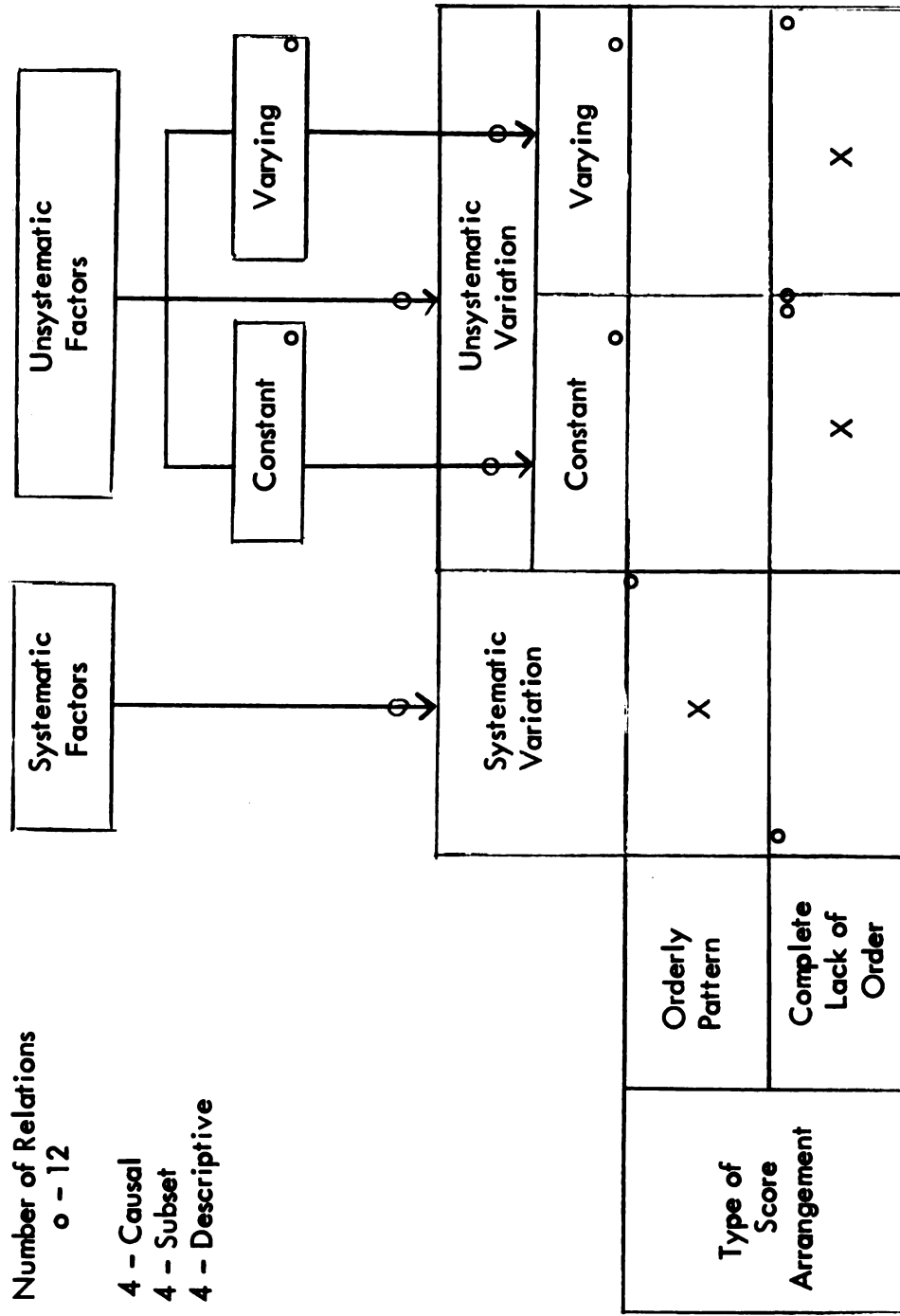
Varying unsystematic factors have different effects on different individuals on the same occasion and have different effects on the same individuals on different occasions. Constant unsystematic factors also have different effects on the same individual on different occasions but have the same effect on all individuals on the same occasion. Systematic factors have the same

effect on all individuals on the same occasion and have the same effect on the same individual on different occasions .

Reliability of measurement is the extent of unsystematic variation in an individual's scores over parallel tests . The reliability coefficient is a quantitative index of this degree of unsystematic variation in an individual's scores over parallel tests . The correlation coefficient is a quantitative index of the degree of unsystematic variation in an individual's scores over parallel and not parallel tests . Parallel tests measure the same trait and meet statistical criteria . Parallel forms of a test measure the same trait and are always similar in content .

The main ways of estimating reliability coefficients are test-retest , parallel forms (delayed and immediate) and internal consistency . Test-retest method involves administering the identical test on different occasions . Parallel forms-delayed method involves administering similar tests on different occasions; parallel forms-immediate method involves administering similar tests on the same occasion . The internal consistency method involves administering the same test once .

APPENDIX G
TEST AND TEST ANALYSIS



General Instructions for the Multiple True- False Questions

In the following questions circle each alternative "true" or "false."

This means that you will mark all alternatives, not just one. Any number of alternatives could be true and any number of them could be false. This, of course, includes the possibility that all could be true or all could be false.

STRUCTURE AND TRANSFER QUESTIONS

1. (STRUCTURE)

If we observed random fluctuation in scores over several occasions we could say that this was caused by

- T F** a. constant unsystematic factors
- T F** b. systematic factors
- T F** c. systematic variation
- T F** d. varying unsystematic factors
- T F** e. constant systematic factors
- T F** f. unsystematic variation
- T F** g. constant systematic variation

Analysis

If alternative a marked true and the rest marked false, then all parts of a must be true: constant unsystematic factors. Total of 4 points.

- (1) causal relationship (CUF-CUV)

(2) subset – factor and variation

(1) random or lack of order description

If alternative a marked true and f also true (rest false), i.e., had both variation and factors as cause, then eliminate constant unsystematic variation descriptive relationship. (3 points)

If alternative d marked true, and the rest marked false, then all parts must be true: varying unsystematic factors. Total of 4 points.

(1) causal relationship (VUF-VUV)

(2) subset – factor and variation

(1) lack of order description

If alternative d marked true and f also true (rest false), then eliminate varying unsystematic variation descriptive relationship. (3 points)

If both a and d marked true and rest false, then an additional 4 points (total of 12)

(1) causal relationship – unsystematic factors

(1) descriptive – random fluctuation – unsystematic variation

(2) can also infer correct relationships for systematic factors and variation (causal and descriptive)

If both a and d are true and f also true (rest false), then eliminate the 3 descriptive relationships for unsystematic variation. (Total of 9 points.)

If a and d are both false, the following possibilities then exist for responding as false.

1. variation instead of factors as cause
2. systematic instead of unsystematic
3. constant and varying describe systematic not unsystematic

If the reason is variation instead of factors, alternatives c and f check this.

- d. systematic variation
- f. unsystematic variation

No points for either of them marked true (and rest false) because cause in wrong direction. Cannot infer correct descriptive relationship either, i.e., might have thought was factor.

If the reason is systematic, alternatives b and c check this.

- b. systematic factors
- c. systematic variation

If b marked true (c true or false, rest false) then had causal relation (factors cause variation) but wrong description. Therefore assume that the subject knew systematic factors cause systematic variation and unsystematic factors cause unsystematic variation. (2 points) If c marked true (d true or false, rest false) then causal relationship incorrect and also descriptive relationship incorrect. (no points)

If the reason is that constant and varying describe (are subsets of) systematic rather than unsystematic, alternatives e and g are pertinent.

- e. constant systematic factors
- g. constant systematic variation

If e marked true (b true or false, rest false) assumed subject knew that factors cause variation (S and U). 2 points

If g marked true (c true or false, rest false) causal and descriptive relationships incorrect. (no points)

Any pattern having both variation and factors as cause and not involving

If a marked true (2 points)

(1) cause

(1) description

If b false (1 point)

(1) definition of systematic variation

Alternative c (Transfer) – on effects of factors. 1 point if marked true.

If d false (2 points)

(1) knew Bill's scores caused by unsystematic factors

(1) description of unsystematic variation

3. (TRANSFER)

If both systematic and unsystematic variation occurred simultaneously within a given set of scores the result would be

T F a. a decrease then an increase in the scores

T F b. random fluctuation only in scores

T F c. an overall trend in the scores with random fluctuations about it.

T F d. lack of order followed by a regular order in the scores

Analysis

If correct combination, then 3 points: 2 for effects and 1 for integration.

If a true, then 2 effects, but wrong combination (2 points)

If b true, then only one effect (1 point)

If d true, then 2 effects, but wrong combination (2 points)

a		T			
b			T		
c	T				
d				T	
	3	2	1	2	0

(Remaining patterns, - 1 point)

4. (STRUCTURE)

John and Bill were both administered a general science test two times. The second administration of the test followed the first by a period of four weeks. On the first administration John had a cold, while on the second he was healthy. Bill was in good health both times. On the first administration of the test the time limit was cut short by an emergency fire drill but on the second administration the time limit was exactly in correspondence with the instructions. Before the first test and also between the two test periods both John and Bill were students in the same general science class. On the first administration Bill had just failed an English test and was quite unhappy. On the second administration John had won a track race and was in good spirits.

In this situation the following factors - health, test time limits, exposure to general science, and mood - affected the test situation and presumably the test results. Classify each of the four situational factors as

S - Systematic factor or

VU - Varying unsystematic factor or

CU - Constant unsystematic factor

Number of Relations (o) - 18

Total of 6 points
(Pattern consists of 3 relationships)

	Effect		Individual(s)			Occasion(s)	
	Same	Different	Same	Different	All	Same	Different
Systematic Factors	o		o				o
	X		X				X
	X	o			o	X	
Unsystematic Factors	X	o			o	o	
		X	oo		X	X	
		oo	oo				oo
Constant		X	X				X
Varying		o		o			

using the above symbols. If the described classroom situation is such that you cannot classify a factor uniquely, list each of the possibilities.

<u>Health</u>	<u>Test Time Limits</u>	<u>Exposure to General Science</u>	<u>Mood</u>
VU	CU	S	VU

5. (STRUCTURE)

A group intelligence test was given to the sixth grade class. The supervising teacher allowed 20 additional minutes for the test. Mary was sitting next to the window and Jane in the dark corner of the room. Different forms of the test were given to Mary and Jane.

Classify these situational factors - time limits, light, and test forms as a specific type of factor, using the following notations:

S - Systematic factor

VU - Varying unsystematic factor

CU - Constant unsystematic factor

If the described classroom situation is such that you cannot classify a factor uniquely list each of the possibilities.

<u>Time Limits</u>	<u>Light</u>	<u>Test Forms</u>
CU, S	VU	VU

Analysis

One of the primary differences between questions 4 and 5 was the time element. Question 4 involved testing on two different occasions while

question 5 involved only one occasion. In each case a correct answer implied a correct pattern of checks (refer to the diagram) across a given row. One point was given for this pattern.

Question 4 - 4 points

Question 5 - 4 points

6. (STRUCTURE)

Below are several statements which may or may not apply to constant unsystematic, varying unsystematic, and systematic factors. If a statement describes a factor mark it accordingly with the following notation: S-Systematic, CU - Constant unsystematic and VU - varying unsystematic. Any statement may be characteristic of more than one type of factor or may not be characteristic of any. So for any statement there is a possibility of 0 to 3 correct answers. If you do not think any factor applies, mark an "X" on the line instead.

<u>S</u>	a. has same effects on same individual on different occasions
<u>VU</u>	b. has effects which can be detected on one occasion
<u>CU, S</u>	c. has same effects on all individuals on one occasion
<u>S</u>	d. produces a decreasing pattern of scores over occasions
<u>VU</u>	e. has different effects on different individuals on the same occasion
<u>CU, VU</u>	f. has different effects on same individual over several testings.

Analysis

One point given for each correct answer (indicated correct pattern).
 One point given for each correct omission (indicated correct pattern).
 Minus one point given for an "X."
 Total of 18 points.

Systematic Factors		X ^o			X ^o
Unsystematic Factors	Const.	X ^o			X ^o
	Vary.	X ^o	X ^o	X ^o	X ^o
		Test Retest	Internal Consistency	Parallel Forms Immediate	Parallel Forms Delayed

This is the transfer structure referred to in the test analysis. Number of relationships - 8.

		T-R	IC	PF-I	PF-D
Type of Test Admin.	Identical	X ^o	X ^o		
	Similar			X ^o	X ^o
Time of Testing	Same Occasion		X ^o	X ^o	
	Different Occasion	X ^o			X ^o
Number of Times Test Admin.	Once		X ^o		
	More Than Once	X ^o		X ^o	X ^o

Number of relationships - 12.

7. (First part - STRUCTURE, Second part - TRANSFER)

The following methods are sometimes used for estimating reliability coefficients:

- (a) test-retest
- (b) internal consistency
- (c) parallel forms - immediate
- (d) parallel forms - delayed

For each of the following situations determine which reliability method is described and place that letter (a,b,c,d) on the line preceding the

corresponding statement. Use only one letter for each statement.

- b A group form of the Stanford-Binet intelligence test was given to the sixth grade class on the opening day of school.

 VU

- d The Iowa Tests of Basic Skills were administered to all transfer students; Form A on the first day of school and Form B two weeks later.

 VU, S, CU

- a The teacher gave the same pre and post test on one chapter in the testbook.

 VU, S, CU

- c Both forms of a personality scale were administered to a group of nurses upon their graduation.

 VU

- b A final examination was given to all students during the last class period of the day.

 VU

You will now be asked to do something else with the same group of statements. Classify each statement as to the type of types of score variation that can be distinguished in each situation. Use this notation

S - Systematic variation

CU - Constant unsystematic variation

VU - Varying unsystematic variation

Use the lines placed below each statement to answer this part of the question.

Analysis

First Part: Description of methods

Each correct answer reflected correct time, type of test and number of testings relationship (3 points). Total - 15 points.

Second Part: Variation and Methods

Each correct answer reflected correct pattern of checks (1 point)

Total - 9 points.

Each correct omission reflected correct pattern of checks (1 point)

Total of 6 points.

Total of 15 points for the transfer part.

8. (TRANSFER)

The four empirical methods of estimating test reliability coefficients are test-retest, internal consistency, and parallel forms, immediate and delayed. You will be asked to list which types of factors affect the reliability coefficients estimated from each of these methods. Use this notation:

S - Systematic factor

VU - Varying unsystematic factor

CU - Constant unsystematic factor.

List the appropriate factor or factors below each method.

<u>Test Retest</u>	<u>Internal Consistency</u>	<u>Parallel Forms Immediate</u>	<u>Parallel Forms Delayed</u>
CU, VU	VU	VU	CU, VU

Analysis

This item checked which factors affect each method as well as the definition of reliability (eliminates systematic factors). Total of 12 points:

6 for each correct unsystematic factor

6 for each omission of systematic factor

9. (TRANSFER)

Suppose we want to determine the reliability of a newly constructed test called the "Teacher Likeability Test." Two parallel forms were prepared. The decision was made to use all four basic methods of estimating reliability coefficients rather than just one or two.

Mark each of the following statements true or false. Assume that similar teachers and group testing procedures were used throughout.

(Explanation of notation used in the statements.)

TR Test-retest

IC Internal Consistency

PFI Parallel forms immediate

PFD Parallel forms delayed

T F a. IC and TR would give the lowest reliability indices.

T F b. PFD and PFI would give the lowest reliability indices.

T F c. TR and PFD would give the highest reliability indices.

Analysis

This item was based on the time of testing as it relates to the distinguishable effects of various factors and to the methods of estimation as well as the definition of reliability coefficient.

Total of 10 points: 6 for constant and varying unsystematic variation operating for different estimation methods.

a		T		
b			T	
c				T
	10	0	0	0

4 for description of test - time of testing for the four methods.

10. (TRANSFER)

T F Assuming parallel tests were used, each method of estimating the reliability coefficient would yield the same reliability coefficient.

Analysis

6 points (How types of unsystematic variation affect methods of estimation.)

11. (TRANSFER)

Theoretically defining reliability in terms of unsystematic variation in scores puts us in a dilemma because in the practical group testing situation

T F a. either only constant unsystematic factors or both varying and constant unsystematic factors can be detected.

T F b. either only varying unsystematic factors or both varying and constant unsystematic factors can be detected.

T F c. either only constant unsystematic factors or only varying unsystematic factors can be detected.

Analysis

If a false and b true, 6 points. Correct relationship between un-systematic factors and estimation methods.

If c false, 1 point. Definition of reliability coefficient.

If a and b false and c true, 0 points. Incorrect relationships, but not inconsistent.

a		T		
b	T			
c			T	
	7	1	0	6

12. (TRANSFER)

If you were attempting to find the reliability of a test, what would you do to improve the estimate of the reliability coefficient?

T F a. individual testing instead of group testing.

T F b. exact timing devices rather than a human being reading a watch

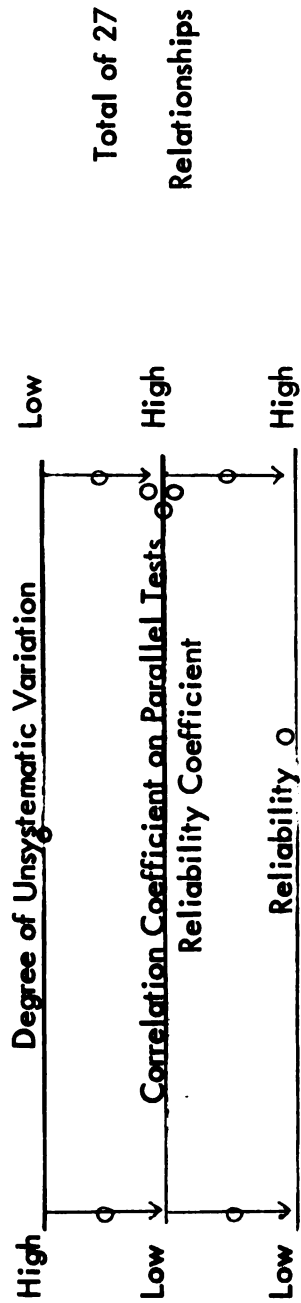
T F c. construct clearer test items.

Analysis

If a false, 1 point. Constant unsystematic variation not distinguishable on one occasion.

If b true, 1 point. Identification of unsystematic variation, reduction of which would improve reliability.

If c true, 1 point. Reduction of unsystematic variation.



9 Relationships

Meet Statistical Criteria	Measure Same Trait	Always Similar in Content
X	X	o
o	X	o

6 Relationships

Tests	Quantitative Index		Reliability Coefficient	Correlation Coefficient
	Degree of Unsystematic Variation in Individual's Scores	Parallel	Not Parallel	
		o	o	o
		X	X	X
		o	o	o

12 Relationships

The following data pertained to the next three questions (13,14 and 15).

The sets of scores over three parallel tests taken by 12 persons are given below.

<u>Person</u>	T E S T S		
	<u>A</u>	<u>B</u>	<u>C</u>
AA	20	25	37
BB	22	26	12
CC	24	26	15
DD	28	28	15
EE	30	29	21
FF	35	30	31
GG	37	31	27
HH	39	32	37
II	40	33	16
JJ	42	34	10
KK	44	35	22
LL	48	36	14

13. (STRUCTURE)

A low relationship between scores on any pair of the tests can be due to

- T F a. a great amount of unsystematic variation in scores
- T F b. the tests not measuring what they claim to measure
- T F c. low correlation coefficient between the tests
- T F d. high correlation coefficient between the tests

Analysis

Alternative a and b were filler items and not scored.

If both c and d false, 3 points.

- (1) correlation not a cause
- (2) infer correct continuums (correlation and unsystematic variation)

If c false and d true, no points. Cause wrong and continuum wrong.

If c true and d false, 2 points. Cause wrong but infer continuums correct.

If both c and d true, then inconsistent, score -1 point.

c			T	T
d		T		T
	3	0	2	-1

14. (STRUCTURE)

If we found a low degree of relationship between test B and C (are parallel tests) we would expect this to be reflected in quantitative indices (index) such as

- T F a. low correlation coefficient
- T F b. high correlation coefficient
- T F c. low reliability
- T F d. high reliability coefficient
- T F e. high reliability
- T F f. low reliability coefficient

Analysis

If a true (b must be false for consistency) - 3 points

(2) quantitative index

(2) two continuums; unsystematic variation and correlation coefficient

If f true (d must be false for consistency) - 3 points

(1) quantitative index

(2) two continuums, unsystematic variation and reliability coefficient

If both a and f true - 1 point.

equality of correlation coefficient on parallel tests and reliability

coefficient.

Consider c and e combinations when both a and f are true:

If both c and e are false - 3 points

(1) reliability not a quantitative index

(2) infer continuums correct - reliability

If c true and e false - 2 points

(2) continuum correct, but quantitative index wrong

If c and e both true - inconsistent

If c false and e true - no points (not a direct contradiction)

If a marked true (b false for consistency) and f false:

f could be false because of two reasons - continuum or quantitative

index. Check this by looking at alternative d (high reliability coefficient)

If d marked true is because of continuum, but incorrect, yet has quantitative index - 1 point.

If d marked false - no points (can not infer about continuum and quantitative index is wrong)

Consider c and e combinations

If c true and e false - 2 points for continuum.

If c false and e false - 1 point for index

If c false and e true - no points .

If a marked false and f marked true (d false for consistency), a could be false because of either continuum or index. Check this by looking at alternative b (high correlation coefficient).

If b true - 1 point for index (continuum wrong)

If b false - no points. (Cannot infer about continuum and index wrong.)

Consider c and e combinations

If c true and e false – 2 points for continuum

If $\bar{c}$ false and $\bar{e}$ false - 1 point for index

If $\bar{c}$ false and $\bar{e}$ true - no points

If both a and f marked false, reason is either continuum or quantitative index. Check by alternative d (high reliability coefficient) and b (high correlation coefficient).

If d true - 1 point for index

If $\bar{d}$ false - no points

If $\bar{b}$ true - 1 point for index

If $\bar{b}$ false - no points

c and e combinations follow same pattern as outlined in the previous section.

[illegible]

15. (STRUCTURE)

The reliability coefficient between tests A and B could be considered a type of correlation coefficient because

- T F a. both coefficients reflect the degree of unsystematic variation in scores
- T F b. both coefficients measure the degree to which tests are always similar in content
- T F c. A and B are parallel, rather than non-parallel tests

Analysis

Alternative b was a filler item.

If a true, 2 points for unsystematic variation in reliability and correlation coefficients.

If c true, 4 points

(2) parallel applied to both

(2) infer correct non-parallel pattern

16. (STRUCTURE)

- T F The degree of unsystematic variation between any two tests would be called reliability (reliability of measurement).

Analysis

If false, 2 points for correct pattern on parallel and non-parallel tests.

17. (TRANSFER)

- T F Parallel tests are harder to construct than parallel forms of a test.

Analysis

If true, 1 point for statistical criteria.

18. (STRUCTURE)

T F Reliability coefficients between parallel tests are high.

Analysis

If false, 1 point for definition of reliability coefficient (UV).

19. (STRUCTURE)

Below are two lists, one of concepts and the other of possible characteristics of those concepts. Before each concept list the characteristics which describe it (use the letter preceding the characteristic.). There may be several describing characteristics or few characteristics for a certain concept.

	<u>Concept</u>	<u>Characteristics</u>
<u>a,b</u>	parallel tests	a. meets statistical criteria
<u>b,c</u>	parallel forms	b. measures same trait
<u>f</u>	reliability	c. similar in content
		d. unaffected by unsystematic factors
		e. constructed from definition of a trait
		f. ranges from high to low

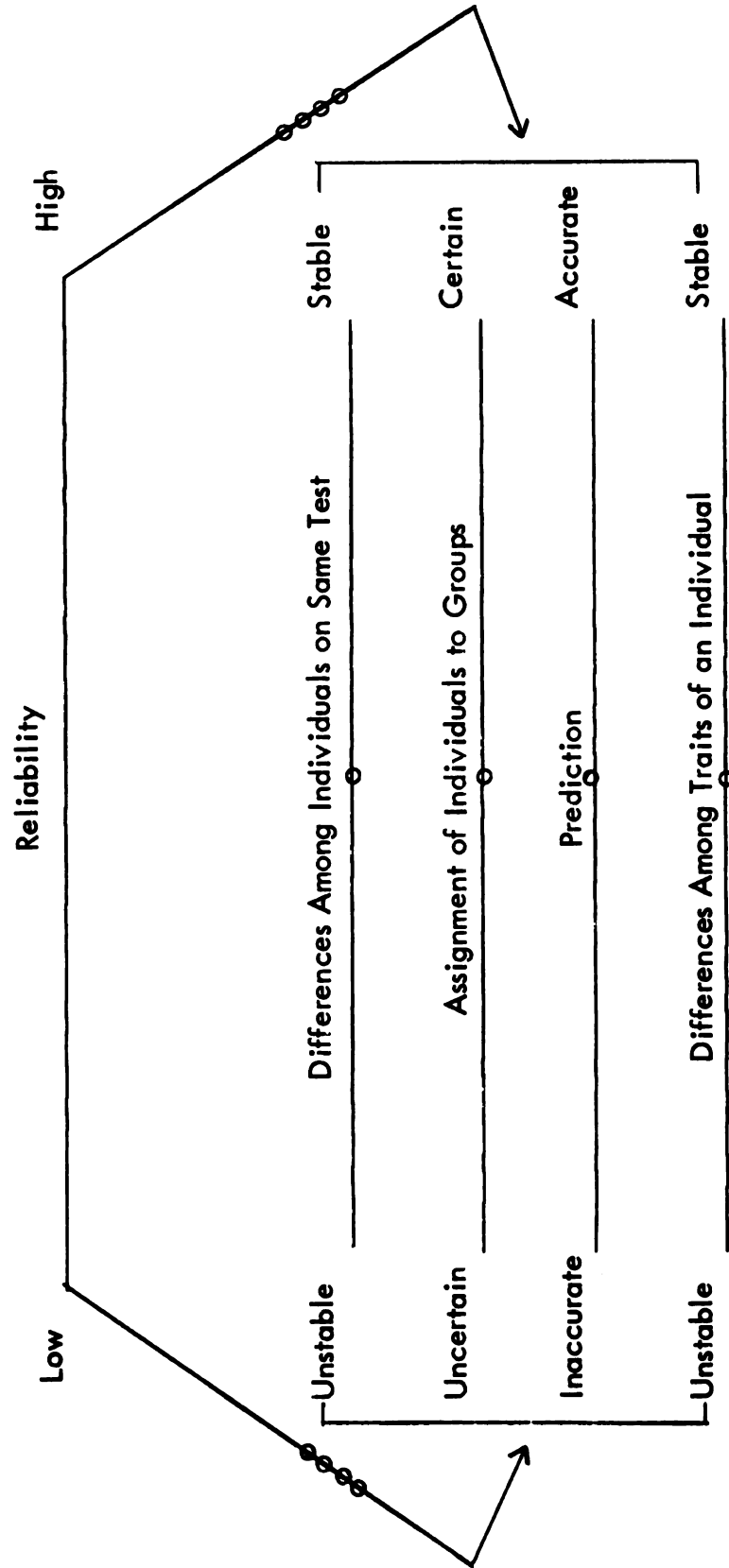
Analysis

Parallel tests

If marked a and b, 2 points for two characteristics
Omission of c, 1 point

Parallel forms

If marked b and a, 2 points for two characteristics



12 Relationships

Omission of a, 1 point

Reliability

If marked f, 1 point for continuum.

Omission of d, 1 point for definition of reliability.

20. (STRUCTURE)

Suppose individuals in one space agency were using psychological tests to screen astronauts for claustrophobia. The tests were reliable. Would you recommend the tests for future use? Answer yes or no.

Analysis

If marked yes, 6 points. (Assignment of individuals to groups and differences among individuals on same test.)

(2) continuums

(4) reliability connection

21. (STRUCTURE)

T F Unsystematic variation in scores from different tests on different traits would be desirable if we were constructing a set of tests for the kit entitled "Guess who is like you." (The purpose is to make a difficult game.)

Analysis

If marked true, 3 points (Differences among traits)

(1) continuum

(2) reliability connection

22. (First part - STRUCTURE, Second part - TRANSFER)

If we wanted to predict how a student taking test X would do on

a biology test, we would prefer that

 T F a. test X be reliable

 T F b. the biology test be reliable

Analysis

If a true, 3 points for continuum and reliability connection.

If b true (Transfer), 3 points for continuum and reliability connection.

ACHIEVEMENT QUESTIONS

Structure Relationship Questions

1. A test is said to be reliable when it
 - a. is published by a reputable company
 - b. provides a basis for diagnosing pupil weaknesses
 - c. can be scored quite easily
 - d. measures what it was designed to measure
 - e. gives an accurate estimate of whatever it measures*
2. Even unreliable scores can be useful to us under the following circumstances:
 - a. comparing different traits of an individual
 - b. comparing individuals on the same list
 - c. predicting behavior
 - d. none of the above*
 - e. all of the above
3. Systematic variation in scores refers to
 - a. a "systematic" distribution of scores in a class (e.g., a normal distribution)
 - b. unbiasedness (e.g., as in a fair dice)
 - c. an orderly sequence of scores*
 - d. none of the above

4. Reliability is a function of
 - a. controlled variation
 - b. systematic variation
 - c. unsystematic variation*
 - d. randomness
5. Tests K and L are parallel tests. In a certain group they correlated .95 and in another the correlation was .20. Such a situation is
 - a. possible, though not common*
 - b. possible, and reasonably common
 - c. mathematically impossible
 - d. impossible by the definition of parallel tests
 - e. impossible, but not for the above reasons
6. The term, varying unsystematic factors, refers to those elements which cause
 - a. variation between individuals in the same situation
 - b. variation within individuals over time, differentially affecting each person
 - c. variation over time, affecting everyone in the group the same
 - d. both a and b*
7. An example of a source of varying unsystematic variation would be
 - a. the test items
 - b. the testee*
 - c. the authors of the test
 - d. the subject matter of the test
8. Contained in constant unsystematic variation would be
 - a. variation between individuals in the same situation
 - b. variation within individuals over time, differentially affecting each person*
 - c. variation within individuals over time, affecting each person the same each time
 - d. both a and b

9. To distinguish between systematic and unsystematic factors we would need to administer a test to
- a. only one person on only one occasion
 - b. only one person on several occasions
 - c. several people on only one occasion
 - d. several people on several occasions*
 - e. in actual practice, it is impossible to distinguish them
10. In categorizing factors, the constant factors are
- a. always placed with systematic factors
 - b. always placed with unsystematic factors*
 - c. placed with systematic or unsystematic, depending on the nature of the factors
 - d. sometimes not placed in either systematic or unsystematic factors
11. A reliability coefficient is obtained by correlating scores on the same form of a test twice administered to the same pupils a number of days apart. Such a reliability coefficient has been termed a
- a. split-half coefficient
 - b. coefficient of equivalence
 - c. internal consistency coefficient
 - d. validity coefficient
 - e. test-retest coefficient
12. How are parallel forms related to parallel tests?
- a. parallel tests are a special type of parallel forms*
 - b. parallel forms are a special type of parallel tests
 - c. they are both the same
 - d. they are actually two rather unrelated terms
 - e. their relationship is more complicated than indicated by any of the above alternatives
13. The computation of internal consistency coefficients requires the administration of
- a. comparable tests to the same group

- b. one test to two groups
- c. comparable tests to different groups
- d. one test to the same group on two occasions
- e. one test to one group*

Related Structure Questions (e.g., may test for presence of elements.)

14. Which of the following is not one of the major types of variances in scores:

- a. constant*
- b. systematic
- c. unsystematic
- d. all of the above
- e. none of the above

15. Systematic variation and systematic factors are terms used by different authors, but refer essentially to the same thing.

T or F

16. Which of the following is an essential concept in reliability theory?

- a. relevance
- b. parallel tests*
- c. parallel forms
- d. criterion measure
- e. all of the above

17. The odd-even method is a special case of the

- a. parallel forms method
- b. internal consistency method*
- c. systematic variation method
- d. test-retest method
- e. none of these

Additional Questions

18. In assigning persons to groups, an unreliable test will likely have the effect of

- a. creating a large middle group
 - b. depleting the middle groups
 - c. increasing the errors of classification*
 - d. none of the above
19. In general, the number of systematic and unsystematic factors which influence a score are approximately:
- a. one
 - b. two
 - c. three
 - d. ten
 - e. none of the above*
20. In which of the following instances would we be most confident of the operating factors?
- a. test scores from groups in two different schools
 - b. test scores from morning and afternoon sessions
 - c. test scores before and after a computer programming course*
 - d. test scores before and after a summer recess
21. The reliability of a reading test of fourth grade pupils is reported to be .78. From this information we can best judge:
- a. how many points pupils are likely to change on the average, if an equivalent test is given
 - b. how many fourth graders are above the norm
 - c. the extent to which each pupil will maintain his position in the group if an equivalent test is given*
 - d. how many fourth graders are below the norm
 - e. the extent to which the test is related to other significant factors in the individual
22. In order to compute a correlation coefficient between traits A and B, it is necessary to have
- a. one group of subjects some of who possess characteristics of trait A, the remainder possess those of trait B
 - b. measures of traits A and B on each subject in one group*
 - c. one group of subjects, some who have both A and B, some with neither, and some with one but not the other

- d. measures of trait A on the group of subjects, and of trait B on another.
 - e. two groups of subjects, one which could be classified as A or not A, the other as B or not B
23. An individual reported a reliability coefficient of an intelligence test as 1.15. It was obtained by correlating the results of a given group on Form A with their results on Form B. This coefficient indicates that
- a. the test has low reliability
 - b. the test is moderately reliable
 - c. the test is highly reliable
 - d. no interpretation can be made without some further crucial information
 - e. a mistake has been made in computing the correlation coefficient*
24. A perfect correspondence or correlation between two variables is represented by a coefficient of
- a. -1.00*
 - b. .00
 - c. .90
 - d. 2.00
 - e. 100.00
25. Which one of these r's has the least predictive value?
- a. .91
 - b. .50
 - c. .17*
 - d. - .23
 - e. -1.00
26. Under a scatter diagram there is a notation that the coefficient of correlation is .06. This means that
- a. most of the cases are plotted within a range of 6% above or below a sloping line in the diagram
 - b. there is a bit more than moderate correlation
 - c. plus and minus 6% from the means includes about 68% of the cases

- d. there is a negligible correlation between the two variables*
 - e. the data mostly (plotted) falls into a narrow band 6% wide.
27. Carry-over effects are most serious with
- a. split half method
 - b. parallel forms method
 - c. test-retest method*
 - d. not very serious with any of the above methods
28. When reliability coefficients can be estimated by several correlation coefficients, one should use the
- a. first one calculated
 - b. median
 - c. arithmetic mean*
 - d. geometric mean
 - e. none of these
29. Internal consistency coefficients are often used because they
- a. are the easiest to compute
 - b. can be calculated from a single administration of a test*
 - c. are easier to interpret
 - d. are the most accurate
 - e. actually, they are seldom used.
30. In determining the quality of a test, reliability is a
- a. desirable but neither a necessary nor sufficient condition
 - b. necessary but not sufficient condition*
 - c. necessary and a sufficient condition
 - d. sufficient but not necessary condition
 - e. none of the above.

APPENDIX H
ORDER OF TEST ITEMS

1.	A12	28.	A19
2.	S1	29.	A3
3	filler item	30.	A22
4.	S2	31.	S11
5.	A1	32.	S22
6.	A27	33.	A21
7	A4	34.	A26
8.	S18	35.	S12
9.	S4	36.	S3
10	A2	37.	A13
11	A5	38.	S6
12	A11	39.	A30
13	A29	40.	S10
14	S5	41	A28
15	A17	42	A15
16	A20	43.	S13
17.	A23	44.	S14
18	S7	45.	S15
19.	A18	46	A10
20.	A24	47.	A7
21	S16	48.	S20
22.	S8	49.	A14
23	S21	50.	S19
24.	A16	51.	A 6
25.	A25	52.	A9
26	S9	53.	A8
27.	S17		

Maximum number of points:

Achievement - 30

Structure - 100

Transfer - 59

APPENDIX I
GUTTMAN DEPENDENCIES

- . A. Questions 4 and 5 jointly dependent upon ability to answer question 6.

All of these questions related to the substructure covering the effects of the three types of factors. Item 6 was a simple recall type item and 4 and 5 were applications of this knowledge. Item 4 pertained to different testing occasions and item 5 pertained to the same testing occasion.

- B Question 7 (second part) dependent upon questions 6 and 7 (first).

Seven-first covered identification of different reliability estimation methods. Seven-second involved listing the type of score variation which could be distinguished in each situation. This required knowledge of the effects of factors on different and same occasions as well as appropriate identification of the type of situation described. The reliability passage did not include the answer to the second part of 7 (transfer). It was expected that both 6 and 7-first would be answered by a great majority of the Ss because these two questions were not very difficult.

- C. Question 8 dependent upon 6 and 7-first (7-second) and 15a and/or 19d.

Question 8 (transfer) tested which types of factors affected reliability coefficients estimated by the various methods. Therefore it required knowledge of the definition of reliability coefficients in terms of degree of unsystematic variation (15a and/or 19d), as well as the interrelationships between factors and estimation methods.

- D. Question 9 dependent upon 6 and 7-first (7-second) and 15a and/or 19d.

Question 9 (transfer) was perhaps more difficult than question 8 and could have been interpreted as an inference from 8 itself. It asked which parts of estimation methods gave the highest and lowest reliability indices. The same reasoning as given in "C" applied here.

- E. Question 11 dependent upon 6 and 7-first (7-second) and 15a and/or 19d.

Question 11 (transfer) set forth a dilemma posed by the differences between the theoretical definition of reliability and methods of estimating reliability coefficients. The same dependency argument used in "C" applied here. Ss were apt to get this item correct by chance partly because of its phrasing.

APPENDIX J
ANALYSIS OF VARIANCE
FOR TIME AND ERRORS

Table 28
Analysis of Variance for
Time and Errors

	Source	SS	df	MS	F
Time					
Group O	Treatment	3373.312	2	1686.66	26.60**
	Error	7101.187	110	63.40	
	Total	10474.499	112		
Group R	Treatment	225.738	2	112.87	1.29
	Error	3499.891	40	87.50	
	Total	3725.629	42		
Groups O and R	Treatment	3801.812	5	760.36	10.90***
	Error	10601.125	150	69.74	
	Total	14402.937	155		
Errors					
Groups O and R	Treatment	5.502	5	1.10	1.41
	Error	117.338	150	.78	
	Total	122.840	155		

*** $p < .001$

** $p < .01$

Table 29
Scheffé Multiple Comparisons on Time

Group O						
	D	V	NR			
(M)	46.53	41.38	33.22			
(SD)	8.57	8.98	5.41			
NR	13.31***	8.159***	. . .			
V	5.151	. . .				
Groups O and R						
	O-D	O-V	R-D	R-NR	R-V	O-NR
(M)	46.53	41.38	40.88	36.36	35.92	33.22
(SD)	8.57	8.98	10.99	8.81	6.07	5.41
O-NR	13.31***	8.16***	7.66*	3.14	2.71	. . .
R-V	10.60***	5.45	4.95	.43	. . .	
R-NR	10.17***	5.02	4.52	. . .		
R-D	5.65	.50	. . .			
O-V	5.15	. . .				

*** $p < .001$

** $p < .05$

APPENDIX K
ANALYSIS OF VARIANCE AND COVARIANCE
FOR ACHIEVEMENT

Table 30
Analysis of Variance for Achievement

		Source	SS	df	MS	F
A1						
Group O	Treatment	1.652	2	.83	.06	
	Error	1498.922	110	13.38		
	Total	1500.574	112			
A1						
Group R	Treatment	4.563	2	2.28	.17	
	Error	533.207	40	13.33		
	Total	537.770	42			
A1						
Groups O R and C	Treatment	1200.05	6	200.01	15.98**	
	Error	2840.91	227	12.52		
	Total	4040.96	233			
A2						
Group O	Treatment	5.301	2	2.65	.20	
	Error	1332.914	100	13.33		
	Total	1338.215	102			
A2						
Group R	Treatment	27.339	2	13.67	.88	
	Error	560.250	36	15.56		
	Total	587.589	38			
A2						
Groups O, R and C	Treatment	914.246	6	152.37	11.80***	
	Error	2749.843	213	12.91		
	Total	3664.089	219			

*** $p < .001$ ** $p < .01$

Table 31
Analysis of Covariance for Achievement

	Source	Adj. MS	df	F
A1, Tm Covariate				
Group O	Treatment	2.235	2	.16
	Error	13.415	109	
A1, Tm Covariate				
Group R	Treatment	1.290	2	.09
	Error	13.577	39	
A2, A1 Covariate				
Group O	Treatment	2.677	2	.36
	Error	7.412	99	
A2, A1 Covariate				
Group R	Treatment	5.457	2	.67
	Error	8.103	35	

Table 32
Scheffé Multiple Comparisons on Achievement
for Six Treatments and Control

A1							
	R-D	R-NR	R-V	O-V	O-NR	O-D	C
(M)	17.06	16.50	16.31	15.53	15.38	15.37	11.06
(SD)	3.36	3.88	3.29	3.67	3.23	3.81	3.31
C	5.99**	5.44**	5.24**	4.46**	4.31**	4.30**	. . .
O-D	1.69	1.13	.94	.16	.01	. . .	
O-NR	1.68	1.12	.93	.15	. . .		
O-V	1.54	.97	.78	. . .			
R-V	.76	.19	. . .				
R-NR	.56	. . .					
A2							
	R-D	R-V	O-D	R-NR	O-NR	O-V	C
(M)	16.86	15.50	15.23	15.00	14.77	14.74	11.06
(SD)	3.08	4.67	3.27	3.82	3.89	3.61	3.31
C	5.81**	4.44**	4.16*	3.94*	3.70*	3.67*	. . .
O-V	2.14	.76	.49	.26	.03	. . .	
O-NR	2.11	.73	.46	.23	. . .		
R-NR	1.88	.50	.23	. . .			
O-D	1.65	.27	. . .				
R-V	1.37	. . .					

** $p < .01$

* $p < .05$

APPENDIX L
ANALYSIS OF VARIANCE AND COVARIANCE
FOR STRUCTURE

Table 33

Analysis of Variance for Structure

Source		SS	df	MS	F
S1					
Group O	Treatment	37.875	2	18.95	.15
	Error	13871.125	110	123.85	
	Total	13909.000	112		
S1					
Group R	Treatment	136.125	2	68.06	.65
	Error	4220.625	40	105.52	
	Total	4356.750	42		
S1					
Groups O, R and C	Treatment	6244.06	6	1040.68	8.09***
	Error	29187.81	227	128.58	
	Total	35431.87	233		
S2					
Group O	Treatment	180.500	2	90.25	1.02
	Error	8864.937	100	88.65	
	Total	9045.437	102		
S2					
Group R	Treatment	140.438	2	70.22	.95
	Error	2664.375	36	74.34	
	Total	2804.813	38		
S2					
Groups O, R and C	Treatment	5927.500	6	987.92	9.00***
	Error	23387.437	213	109.80	
	Total	29314.937	219		

*** p < .001

Table 34
Analysis of Covariance for Structure

	Source	Adj. MS	df	F
S1, Tm Covariate				
Group O	Treatment	11.732	2	.09
	Error	122.416	109	
S1, Tm Covariate				
Group R	Treatment	63.089	2	.58
	Error	107.959	39	
S2, S1 Covariate				
Group O	Treatment	72.669	2	.13
	Error	57.754	99	
S2, S1 Covariate				
Group R	Treatment	67.398	2	.10
	Error	65.839	35	

Table 35

Scheffé Multiple Comparisons on Structure
for Six Treatments and Control

S1							
	R-NR	R-D	O-D	O-V	O-NR	R-V	C
(M)	64.93	64.50	62.74	62.47	61.32	60.85	51.81
(SD)	7.99	10.73	11.64	8.95	11.52	11.35	12.31
C	13.12**	12.69*	10.93*	10.67*	9.52*	9.04	. . .
R-V	4.08	3.65	1.89	1.63	.48	. . .	
O-NR	3.60	3.18	1.41	1.15	. . .		
O-V	2.46	2.03	.26	. . .			
O-D	2.19	1.76	. . .				
R-D	.43	. . .					
S2							
	R-NR	O-V	R-D	R-V	O-D	O-NR	C
(M)	66.46	63.38	63.25	61.70	61.09	60.24	51.81
(SD)	7.24	8.15	7.83	10.01	9.78	9.79	12.33
C	14.64**	11.58**	11.44**	9.89	9.28**	8.43*	. . .
O-NR	6.23	3.15	3.02	1.47	.85	. . .	
O-D	5.33	2.30	2.16	.61	. . .		
R-V	4.76	1.68	1.55	. . .			
R-D	3.21	.13	. . .				
O-V	3.08	. . .					

*** $p < .001$

** $p < .01$

* $p < .05$

APPENDIX M
ANALYSIS OF VARIANCE AND COVARIANCE
FOR TRANSFER

Table 36
Analysis of Variance for Transfer

Source		SS	df	MS	F
T1					
Group O	Treatment	128.000	2	64.00	.86
	Error	8325.000	110	74.34	
	Total	8454.000	112		
T1					
Group R	Treatment	42.996	2	21.50	.35
	Error	2425.984	40	60.65	
	Total	2468.980	42		
T1					
Groups O, R and C	Treatment	522.812	6	87.14	1.34
	Error	14735.062	227	64.91	
	Total	15257.874	233		
T2					
Group O	Treatment	64.160	2	32.08	.60
	Error	5313.687	100	53.14	
	Total	5377.847	102		
T2					
Group R	Treatment	30.805	2	15.40	.23
	Error	2419.965	36	67.22	
	Total	2450.770	38		
T2					
Groups O, R and C	Treatment	759.563	6	126.59	2.40*
	Error	11216.562	213	52.66	
	Total	11976.125	219		

* $p < .05$

Table 37
Analysis of Covariance for Transfer

	Source	Adj. MS	df	F
T1, Tm Covariate				
Group O	Treatment	39.519	2	.53
	Error	74.746	109	
T1, Tm Covariate				
Group R	Treatment	20.546	2	.38
	Error	53.907	39	
T2, T1 Covariate				
Group O	Treatment	16.553	2	.38
	Error	43.993	99	
T2, T1 Covariate				
Group R	Treatment	23.541	2	.57
	Error	41.011	35	

APPENDIX N
ANALYSIS OF VARIANCE FOR
SUBSTRUCTURES

Table 38

Analysis of Variance for Substructures on Acquisition -
Six Treatments and Control

	Source	SS	df	MS	F
Sb1	Treatment	443.101	6	73.85	3.37**
	Error	4978.062	227	21.93	
	Total	5421.163	233		
Sb2	Treatment	328.551	6	54.76	3.70**
	Error	3356.414	227	14.79	
	Total	3684.965	233		
Sb3	Treatment	350.660	6	58.44	4.47***
	Error	2969.957	227	13.08	
	Total	3320.617	233		
Sb4	Treatment	607.438	6	101.24	4.57***
	Error	5025.664	227	22.14	
	Total	5633.102	233		
Sb5	Treatment	111.941	6	18.66	9.48***
	Error	446.674	227	1.97	
	Total	558.615	233		
Sb6	Treatment	10.875	6	1.81	.34
	Error	1214.511	227	5.35	
	Total	1225.386	233		

*** $p < .001$

** $p < .01$

Table 39
Analysis of Variance for Substructures
on Retention - Six Treatments

	Source	SS	df	MS	F
Sb 1	Treatment	117.543	5	23.51	1.21
	Error	2632.887	136	19.36	
	Total	2750.430	141		
Sb 2	Treatment	25.391	5	5.08	.40
	Error	1727.598	136	12.70	
	Total	1852.989	141		
Sb 3	Treatment	92.195	5	18.44	1.72
	Error	1461.382	136	10.75	
	Total	1553.577	141		
Sb 4	Treatment	79.973	5	15.99	.90
	Error	2423.324	136	17.82	
	Total	2503.297	141		
Sb 5	Treatment	6.443	5	1.29	.68
	Error	256.775	136	1.89	
	Total	263.218	141		
Sb 6	Treatment	22.379	5	4.48	1.12
	Error	545.320	136	4.01	
	Total	567.699	141		

Table 40

Scheffé Multiple Comparisons on Substructure -
Acquisition for Six Treatments and Control

Sb1	O-D	O-NR	R-D	R-V	O-V	R-NR	C
(M)	9.05	8.38	7.75	7.31	7.11	6.07	5.44
(SD)	4.49	4.93	5.55	4.93	4.83	4.74	4.81
C	3.62*	2.94*	2.31	1.87	1.67	.64	...
R-NR	2.98	2.31	1.68	1.24	1.03	...	
O-V	1.95	1.27	.65	.20	...		
R-V	1.75	1.07	.44	...			
R-D	1.30	.63	...				
O-NR	.67	...					
Sb2	R-NR	R-D	O-NR	O-V	R-V	O-D	C
(M)	17.79	17.06	17.03	16.42	15.54	14.92	14.41
(SD)	1.82	3.25	3.23	3.18	4.34	4.02	4.40
C	3.38	2.65	2.62	2.01	1.13	.41	...
O-D	2.87	2.14	2.11	1.50	.62	...	
R-V	2.25	1.52	1.49	.88	...		
O-V	1.37	.64	.61	...			
O-NR	.76	.04	...				
R-D	.72	...					
Sb3	R-NR	O-V	R-D	O-D	R-V	O-NR	C
(M)	13.29	12.95	12.56	12.24	11.54	10.46	10.12
(SD)	3.15	3.09	2.18	2.87	3.88	4.10	4.02
C	3.17	2.83	2.45	2.12	1.42	.34	...
O-NR	2.83	2.49	2.10	1.78	1.08	...	
R-V	1.75	1.41	1.02	.70	...		
O-D	1.05	.71	.33	...			
R-D	.72	.39	...				
O-V	.34	...					

Table 40-- (Continued)

Sb4	R-NR	R-D	R-V	O-NR	O-V	O-D	C
(M)	17.21	16.44	16.39	16.34	16.16	16.00	12.94
(SD)	2.78	3.86	2.62	4.62	3.31	4.68	5.72
C	4.28	3.50	3.45	3.41	3.22	3.06	. . .
O-D	1.21	.44	.39	.16	.16	. . .	
O-V	1.06	.28	.23	.18	. . .		
O-NR	.87	.09	.04	. . .			
R-V	.83	.05	. . .				
R-D	.78	. . .					
Sb5	O-D	R-D	R-NR	R-V	O-V	O-NR	C
(M)	4.03	3.94	3.93	3.85	3.61	3.29	2.33
(SD)	1.49	1.56	.56	1.29	1.44	1.16	1.46
C	1.69**	1.60*	1.59	1.51	1.27	.97	. . .
O-NR	.73	.64	.63	.55	.31	. . .	
O-V	.42	.33	.32	.24	. . .		
R-V	.18	.09	.08	. . .			
R-NR	.09	.01	. . .				
R-D	.08	. . .					
Sb6	R-D	R-NR	C	O-V	R-V	O-NR	O-D
(M)	6.75	6.64	6.58	6.14	6.23	6.16	6.16
(SD)	2.25	2.02	2.28	2.42	2.75	2.31	1.94

** $p < .01$

* $p < .05$

Table 41

Means and Standard Deviations on Substructure -
Retention for Six Treatments

Sb1	O-NR	O-V	O-D	R-NR	R-V	R-D
(M)	6.71	9.00	7.51	8.92	7.60	8.63
(SD)	3.91	4.39	4.42	4.37	4.48	4.47
Sb2	O-NR	O-V	O-D	R-NR	R-V	R-D
(M)	16.24	16.12	15.63	16.54	15.30	16.81
(SD)	3.39	2.91	4.27	4.38	2.57	2.45
Sb3	O-NR	O-V	O-D	R-NR	R-V	R-D
(M)	11.38	13.24	12.00	13.62	12.60	11.63
(SD)	3.81	2.73	3.44	3.81	2.62	3.81
Sb4	O-NR	O-V	O-D	R-NR	R-V	R-D
(M)	16.15	15.09	15.40	17.39	16.60	16.81
(SD)	4.51	4.21	4.74	2.65	3.44	2.89
Sb5	O-NR	O-V	O-D	R-NR	R-V	R-D
(M)	3.24	3.68	3.34	3.54	3.30	3.00
(SD)	1.06	1.32	1.53	1.08	1.62	1.50
Sb6	O-NR	O-V	O-D	R-NR	R-V	R-D
(M)	6.53	6.27	7.29	6.46	6.30	6.38
(SD)	2.12	1.97	1.79	1.59	2.49	1.79

APPENDIX O

QUESTIONNAIRE ITEMS UNIQUE TO
DIAGRAM AND VERBAL TREATMENTS

Table 42

Responses to the Diagram and Verbal
Questionnaire Items

Question	Response		
	Yes	No	No Response
Diagram Treatment			
1. Examine small diagrams	49	2	1
Trouble with interpretation	11	37	2
2. Examine large diagrams	51		
Trouble with interpretation	24	26	1
Examine inter-connections	47	3	1
Randomly	30		
Systematically	15		
3. (Use of diagrams while reading)			
Repeat	19		
Integrate	18		
Check on learning	19		
Organize	18		
Remember spatially	17		
Other	4		
4. (Use of diagrams during test)			
Visualized diagram	10		
Recognized connection	13		
Vague remembrance	26		
No recall	6		

Table 42-- (Continued)

Question	Yes	Response	
		No	No Response
Verbal Treatment			
1. Examine small reviews	37	2	
2. Read large review	36	3	
3. (Use of review passage while reading)			
Repeat	21		
Integrate	8		
Check on learning	17		
Organize	8		
Remember verbally	13		
Other	1		
4. (Use of review during test)			
Instant recognition	9		
Vague remembrance	25		
No recall	7		

APPENDIX P

APTITUDE CORRELATIONS

Table 43
Correlations among Aptitude Scores
and Main Dependent Variables

	O-V, ACE, n = 11			O-V, CAAT, n = 8		
	Q	V	T	Q	V	T
A1	-.109	.490	.300	.318	.560	.545
A2	-.093	.464	.289	.168	.434	.370
S1	.329	.673*	.611*	.353	.565	.572
S2	.496	.620*	.644*	.315	-.050	.192
T1	.165	.347	.314	-.453	-.244	-.460
T2	.186	.522	.445	-.305	-.344	-.414
T _m	-.128	-.120	-.138	-.724*	.439	-.261
E	.119	.118	.133	-.470	.012	-.326

	O-D, ACE, n = 17			O-D, CAAT, n = 9		
	Q	V	T	Q	V	T
A1	.072	.157	.148	.834**	.212	.620
A2	.531	.527	.434	.614	.407	.499
S1	-.044	.057	.030	.323	.185	.298
S2	.055	.507	.384	.779**	.373	.677*
T1	.432	.394	.502*	.797**	.027	.493
T2	-.037	.170	.102	.207	.455	.380
T _m	-.389	-.076	-.250	-.025	-.374	-.225
E	-.130	.134	.027	-.084	-.078	-.094

R-NR, ACE, n = 7			
A1	-.681	.374	.099
A2	-.823	.044	-.224
S1	-.799	-.232	-.449
S2	-.748	.079	-.171
T1	.556	.664	.737
T2	-.525	.112	-.072
T _m	-.593	-.375	-.505
E	-.792*	-.623	-.778*

** p < .01

* p < .05

MICHIGAN STATE UNIVERSITY LIBRARIES



3 1293 03071 3063