THE GROUP MEASUREMENT OF GENERALIZING ABILITY AT THE GRADE SIX LEVEL

by

Gordon Thomas Filmer-Bennett

A Thesis submitted in Partial Fulfilment of
The Requirements for the Degree of

M A S T E R OF A R T S

in the Department of
Philosophy and Psychology

THE UNIVERSITY OF BRITISH COLUMBIA

October 1946.

ACKNOWLEDGMENT

I wish to extend thanks to Mr. R. Straight, Superintendent of the Vancouver City Schools, and to the principals and teachers who co-operated so willingly in the execution of this project. I am also indebted to Mr. T. Robinson and more especially to Miss T. Combolos for assistance in test administration and scoring. Credit is also due Mr. H. Parker who shared in the construction of these tests,

Accepted Oct 9/46.
130/150

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont'd)

LIST OF TABLES AND PLATES

A. TABLES

LIST OF TABLES AND PLATES (Cont'd.)

LIST OF TABLES AND PLATES (Cont'd.)

LIST OF TABLES AND PLATES (Cont'd.)

B. PLATES

I N T R O D U C T I O N

The demands of a constantly changing world coupled with the increasing emphasis upon scientific techniques bespeak the need for systematic thinking of the highest order. An inherent part of such thinking, the abstraction of meaning and the formation of concepts represent, according to Sherman[1], the acme of intelligence. As he points out, there is general agreement that "an accurate measure of a person's intelligence is possible only when his capacity to form and express concepts (abstract thinking) can be estimated." The fact that conceptual thinking is of such undisputed importance in directing human activity stimulates interest in developing ways and means toward its analysis and measurement.

Since conceptual ability falls within the realm of thinking normally designated as reasoning, the search for adequate definitions might well commence with the latter,

---

[1] Sherman, M., Intelligence And Its Deviations, New York, The Ronald Press Co., 1945, p. 15.

According to Billings[1], reasoning consists in "the solving of a practical or theoretical problem or difficulty by the use of or through the relating of past experiences." Extending this definition, Gates[2] classifies reasoning as a "a form of learning", not vastly unlike trial-and-error learning, in which pertinent facts are recalled and combined with those perceived at the time. The popularity of this view is indicated by the frequent reference to rational learning, concept-learning, and the like in psychological journals and textbooks. While not in fundamental disagreement with this conception, McGeoch[3] prefers to regard learning and reasoning as separate but closely related processes operating from near-opposite ends of a continuum, with overt response in the one giving way to symbolic response in the other. Following upon a series of fact-finding endeavours, an interpretation offered by Maier[4] regards reasoning as spontaneous adjustment, a type of integrative response which depends for success upon the removal of persistent and old-established tendencies. It is the ease and readiness with which these persistencies of habit are sidetracked that distinguishes between able and poor reasoners. Furthermore, reasoning and learning ability are not necessarily highly correlated; a clever reasoner

1. Billinga, M.L., "Problem-Solving In Different Fields of Endeavor", American Journal of Psychology, vol. XLVI,1934, p. 260.
2. Gates, A.I., "Psychology for Students of Education, New York, The MacMillan Co., 1930, pp. 386-393.
3. McGeoch, J.A., Psychology of Human Learning, New York, Longmans, Green and Co., 1942, p. 517.
4. Maier, N.R.F., "Reasoning in Rats and Human Beings", Psychological Review, vol. XLIV, 1937, pp. 365-378.

may be a poor learner, and vice-versa. Learning "furnishes us data with which to solve problems", but it "also furnishes us with habitual directions and so interferes with new adaptations." That is why age and "excess" learning are more often than not productive of stereotyped ways of thinking which becloud the possibility of new approaches toward the solution of a problem. "By regarding reasoning as a new combination of past experiences," Maier concludes, "we designate a mechanism which differs from learning and yet utilizes what has been learned."

On the whole, the apparent controversy over the interrelation of reasoning and learning seems to spring mainly from differences in the definition of learning. As to what constitutes reasoning, there is general acknowledgment of the presence of a problem requiring solution, of the need for recall, and of the importance of past experience and the relating and reorganization of pertinent parts of this experience.

Turning to a quantitative analysis of reasoning, Thurstone[1] found that, instead of being highly specific, reasoning appeared divisible into merely two factors which, for want of further study, he tentatively labelled "I" and "D" Factor "I" is linked to the discovery of a rule or to the

---

1. Thurstone, L.L., "Primary Mental Abilities", Psychometric Monographs, #1, 1938, pp.v, 86-89.

formulation of a hypothesis, and is therefore representative of induction; factor "D" is associated with the application of a general rule or principle to particulars, thus symbolizing deduction. Recent experiments by Holzinger and others have suggested the possibility, however, that these may not actually exist as separate factors at all[1]. Whatever the facts, induction and deduction would seem to be intimately related.

Aware of the probable overlap of inductive and deductive thinking, Woodworth[2] regards these terms as more aptly describing problems than thought processes. He uses "induction" synonymously with "concept formation" in reference to problems which call forth classificatory or generalized responses. Early though not wholly representative examples of this type of problem were those used by Hull and Kuo[3], in which the abstraction of common elements from a series of patterns was considered an aspect of concept formation. In neither of these cases, however, was there provision for generalization in the sense in which Smoke[4]

1. Wolfe, D., "Factor Analysis to 1940", Psychometric Monographs, #3, 1940, p. 33.
2. Woodworth, R.S., Experimental Psychology, London, Methuen and Co. Ltd., 1938, p. 801.
3. Hull, C.L., "Quantitative Aspects of the Evolution of Concepts; An Experimental Study", Psychological Monographs, vol. XXVIII, No.1, 1920; Kuo, Z.Y., "A Behavioristic Experiment on Inductive Inference", Journal of Experimental Psychology, vol. VI, 1923, pp.247-293. Both cited in Smoke, K.L., "An Objective Study of Concept Formation".
4. Smoke, K.L., "An Objective Study of Concept Formation", Psychological Monographs, vol. XLII, No.4,1932, pp.2-8, 42.

employs the word. For Smoke, generalization or concept formation is something more than the mere abstracting of elements; it is a search for common relationships. Response is no longer to a single element within the stimulus pattern, but to a "dynamic whole". Elements, according to Smoke, may not even enter into the concept. But Tyler[1] contends that generalization "involves both elements and relations between these elements". It is indeed difficult to conceive of "relationship" as an entity in itself, for the very term implies "things related". However, this fact does not detract from Smoke's definition of generalization or concept formation as a "process whereby an organism develops a symbolic response (usually but not necessarily linguistic) which is made to the members of a class of stimuli patterns, but not to other stimuli"[2]. Since response in this sense involves the formulation of a rule or principle, generalization or concept formation may be regarded as nothing less than an expression of reasoning ability. This definition of generalization, therefore, will be applied in the present study.

---

1. Tyler, F.T., Generalizing Ability of Junior High School Pupils: An Experimental Study of Rule Induction, unpublished Ph.D. Thesis, University of Californis, 1939.
2. Smoke, K.L., op. cit., p. 8.

CHAPTER I.

THE PROBLEM

1. Review of the Literature.

Hull and Kuo pioneered the way toward an objective analysis of generalizing ability, but it remained for others to expand the technique. An inventory of such experiments to date indicates the development of several sub-types which, for immediate purposes, will be reduced to those emphasizing simple sorting tests and those favouring other means for the study of generalizing ability. Typical of the former are the experiments of Hanfmann and Kasanin.[1] Their tests, administered individually, required the classification of geometric solids according to the possession of certain common properties. They outlined three significant characteristics of conceptual thinking, namely, "the importance of the attitude of looking for categories, the recognition of many possibilities rather than merely the first one to occur, and the consideration of the total system". Conducting an experiment along almost identical lines, Thompson[2] found quantitative and qualitative differences between the generalizing ability of 6- to 8-year-olds and that of 9-to 11-year-olds, the latter exhibiting less

1. Hanfmann, E., & Kasanin, J., "A Method for the Study of Concept Formation", Journal of Psychology, vol. III, 1937, pp. 524-529.
2. Thompson, J., "The Ability of Children of Different Grade Levels to Generalize on Sorting Tests", Journal of Psychology, vol. XI, 1941, pp. 119-126.

rigidity in their attack upon the problems. Her results are closely allied to those of Long and Welch[1] in showing that classification on the basis of form is probably one of the lowest levels of generalization. On the whole, these experiments involved a relatively small number of subjects and were sparing in their use of statistical analysis.

The other type of study is exemplified by the individual experiment of Ewart and Lambert[2] wherein the subject advanced toward a solution of the problem through the perception of a complexity of positional relationships. Generalization was found to be highly correlated with intelligence, and to benefit from verbal instruction. Conclusions were based upon a small, select group and made no reference to reliabilities or sev differences. A group experiment by Peterson[3] required the derivation of a general rule or principle (physical law of the lever) operating in each of a series of 20 problems. Performance was rated according to the number of problems solved and a correct statement of the underlying principle involved. The ability to solve problems in this setting bore little or no relation to intelligence. It also appeared that success in solving the problems was adversely affected by a

1. Long, L., and Welch, L., "A Preliminary Investigation of Some Aspects of the Hierarchical Development of Concepts", Journal of General Psychology, vol. XXII, 1940, pp.359-388
2. Ewart, P.H., & Lambert, J.F., "The Effect of Verbal Instructions Upon the Formation of a Concept", Journal of General Psychology, vol. VI, 1932, pp.400-413.
3. Peterson, G.M., "An Empirical Study of the Ability to Generalize", Journal of General Psychology, vol.VI, 1932, pp.90-114.

reduction in the amount of instruction forthcoming.

Tyler[1] conducted an investigation of generalizing ability, using a combination light and switch panel. The problem was to discover from patterns arranged thereon, the switch which turned out all the lights. This was an individual experiment. Correlations with intelligence were significant and substantial. Sex differences favouring the boys were probably linked to the mechanical nature of the apparatus. Results suggested that solutions were achieved with the aid of both positive and negative instances, where "positive" was used to describe those examples which illustrate the rule governing solution, and "negative" to examples which violate this rule. Tyler also found that solutions were not always accompanied by the ability to verbalize the rule or principle concerned.

Sidestepping the need for overt manipulation, Smoke[2] required his subjects to discern common relationships between elements contained within a series of geometric patterns. As in the preceding experiment, successful generalization did not imply ability to define the concept verbally. The negative teaching example (defined as in Tyler's experiment) promoted greater accuracy, but had a less decided effect upon rate of performance; a majority

---

1. Tyler, F. T., op. cit.
2. Smoke, K.L., op. cit.; Smoke, K.L., "Negative Instances in Concept Learning", Journal of Experimental Psychology, vol. XVI, 1933, pp. 583-8.

preference lay with the use of both positive and negative teaching examples. Among factors observed to be characteristic of concept formation were (1) grouping, (2) insightful behavior, and (3) formulation, testing, and acceptance or rejection of hypotheses. Computing correlations for one experimental group, intelligence and speed of generalizing were found to be significantly related. This relationship was not determined for the other groups, nor were comparisons made with performance on other reasoning tests. As Tyler has previously noted, no study was made of sex differences nor of the effect of order of presentation upon item-difficulty. The subjects were tested individually, and together formed a relatively small and highly select group.

Foreseeing the possibilities behind Smoke's technique, Tyler suggested the need for its further application. To this end Wood[1] made numerous changes in Smoke's tests to permit their adaptation to a lower age level; the method of presentation was altered somewhat and most of the items were completely redefined. The tests were administered individually to 50 Grade VI boys, half of whom were subjected to instruction by means of positive examples, while the remainder were taught by both positive and negative examples. In each case the teaching examples

1. Wood, J.E., The Relative Role of Positive and Negative Instances In Concept Formation, unpublished Master's Thesis, Vancouver, University of British Columbia, 1943.

were presented in cumulative fashion and were allowed to remain before the subjects throughout the testing period. On the basis of his results Wood concluded that the negative teaching example greatly assists generalization, especially among those of lesser intelligence and in those cases involving more complex items. Recognition was a more reliable measure of generalizing ability than verbalization, though no one of recognition, verbalization, or reproduction could be depended upon to precede the others in order of appearance. In spite of the limited size of the groups, correlations with other variables might have been computed. No provision was made for the study of sex differences, nor were reliabilities listed. Performance being rated solely according to the number of perfect scores, there was no differentiation between individuals of unequal ability who were both capable of obtaining the solution to an item. Had a time limit been imposed and all test trials been made compulsory, a composite score made up of perfect scores and number of trials required to reach a solution would have yielded a more accurate measurement of the ability under consideration.

The first to apply Smoke's principle in a group experiment, Dickinson subjected the test items to still further changes and modified procedure in accordance with

1. Dickinson, A.E., An Investigation Into The Generalizing Ability Of Grade Two Pupils; Master's Thesis, Vancouver, University of British Columbia, 1943, published in abstract in Journal of Educational Psychology, vol. XXXV,1944. pp. 432-441.

several of Wood's findings by reducing the number of teaching and test instances. With 160 Grade II children for subjects, she made a comprehensive study of the effect upon generalizing ability of instruction utilizing successive and cumulative presentation of both positive and negative examples and of positive examples only. Teaching examples successively presented were removed during the testing period, hence involving the need of recall; under cumulative presentation they were continuously exposed as in Wood's experiment. Subjects were selected to form four groups of 20 boys and 20 girls each, matched on the basis of intelligence and chronological age. Achievement, registered in terms of mean scores rather than number of perfect scores, was most successful under successive presentation and was more impaired than aided by the introduction of negative teaching examples, though these trends were not statistically significant. Boys showed superior ability when instructed by positive and negative examples, and girls when instructed by only positive examples; again, however, these differences were not dependable. Test reliabilities were high. Correlations with intelligence and reading ability were, in general, low and negligible, but the relationship of test performance to scholastic achievement was not determined. These results should be confirmed by employing larger samples embodying a wider intelligence range.

2. Summary of the Literature

From this condensed account of related studies in generalizing ability emerge the following conclusions:

1. Differences in experimental results, which are probably attributable to group differences as well as to differences in test material and methods of procedure, illustrate the need for more repetition and follow-up of experiments previously undertaken.
2. Wherever possible, individual experiments should be repeated as group experiments, and vice versa.
3. Verbalization of thexrule or principle governing solution of a problem is a doubtful criterion of generalizing ability or concept formation.
4. The question of sex differences and the value of the negative teaching example demand closer study.
5. Consideration of test validity was usually restricted to correlations with various criteria; no reference was made to item validity.
6. No attempt was made to analyze reaction to positive and negative test instances.

3. The Problem Defined

Generalizing ability may be estimated by any one of the experimental methods previously described, but that utilized by Smoke appears most suited to both individual and group testing at any level. Smoke's technique permits

as close an approach to the study of the ordinary everyday process of conceptual thinking as any yet devised. The present problem may be broadly defined as the group measurement of generalizing ability at the Grade VI level, where "generalization" is used synonymously with "concept formation" to designate the process whereby a common relationship is abstracted from a series of geometric patterns. Its basic assumptions are that

(a) Generalizing ability, if possessed by children at the Grade VI level, can be measured by the method described.

(b) Generalizing ability is more accurately represented by scores on a recognition test than by verbalization of the rule involved in the solution.

4. The Problem in Outline

Specific questions which this study will attempt to investigate may be outlined in brief.

1. What is the effect upon generalizing ability of group instruction which calls for the exposure one by one of patterns representative of the rule or principle to be deduced and which requires their removal during the testing period?

2. What is the effect upon generalizing ability of group instruction which calls for the alternate exposure one by one of patterns representative and of patterns not representative of the rule or principle to be deduced and which requires their removal during the testing period?
3. What is the effect upon generalizing ability of group instruction which calls for the alternate exposure by cumulative presentation of patterns representative and of patterns not representative of the rule or principle to be deduced and which permits their continued exposure during the testing period?
4. To what extent, if any, do sex differences govern generalizing ability in this setting?
5. With what reliability and accuracy can the group measurement of generalizing ability at the Grade VI level be accomplished? How closely related to other forms of mental achievement is the ability to abstract spatial relationships?
6. What are some of the factors of difficulty which impede successful generalization of this type?
7. Are test stimuli which exemplify the rule or principle governing solution and those which do not illustrate this rule identified with equal accuracy?

CHAPTER II.

GENERAL PROCEDURE, APPARATUS, AND SUBJECTS

1. General Procedure.

Since this study was conducted with a view to retesting several hypotheses advanced by Smoke and Wood, and to determining the extent to which their results pertain to group situations, it was desirable that the general conditions surrounding concept formation in the present investigation parallel closely those of the previous studies.

The nine different geometric symbols or "concepts" constituting the present tests were borrowed from Wood who, in turn, designed them from Smoke's. A nonsense syllable was used in both cases to designate a given series of geometric patterns exhibiting a common relationship between certain elements contained within them. These "concepts" and their accompanying definitions are listed below in the order in which they were presented to the subjects.

Concept	Definition
Dax:	A triangle containing a dot.
Mef:	A circle, half black and half white.
Vec:	A straight line, at one end of which is axdot in direct line with it.

Concept	Definition
Mib:	A circle touching a square.
Zum:	A circle with one dot inside and one dot outside it.
Tov:	A square and four crosses, one near each of its four sides.
Pog:	Two lines (straight or otherwise) of unequal length.
Wez:	A circle touching the shortest side of a triangle.
Zif:	A circle inside a rectangle, and touching its two longest sides but Not touching either end.

Sets of traching and test instances, identical to those used by Wood, with the exception of a slight change in the order of presentation and in the number of teaching and test instances employed, were prepared. Instead of eight teaching examples as in Wood's experiment four such examples of a given concept were presented. This procedure, already adopted by Dickinson, is in conformity with Wood's findings, namely, that performance showed little or not improvement beyond the fourth presentation. Likewise, the number of test instances was reduced from sixteen to ten. Hereafter the terms "example" and "instance" will be used to distinguish patterns comprising the teaching series and those comprising the test series, respectively. When referring collectively to teaching and test patterns, the term "instances" will be applied.

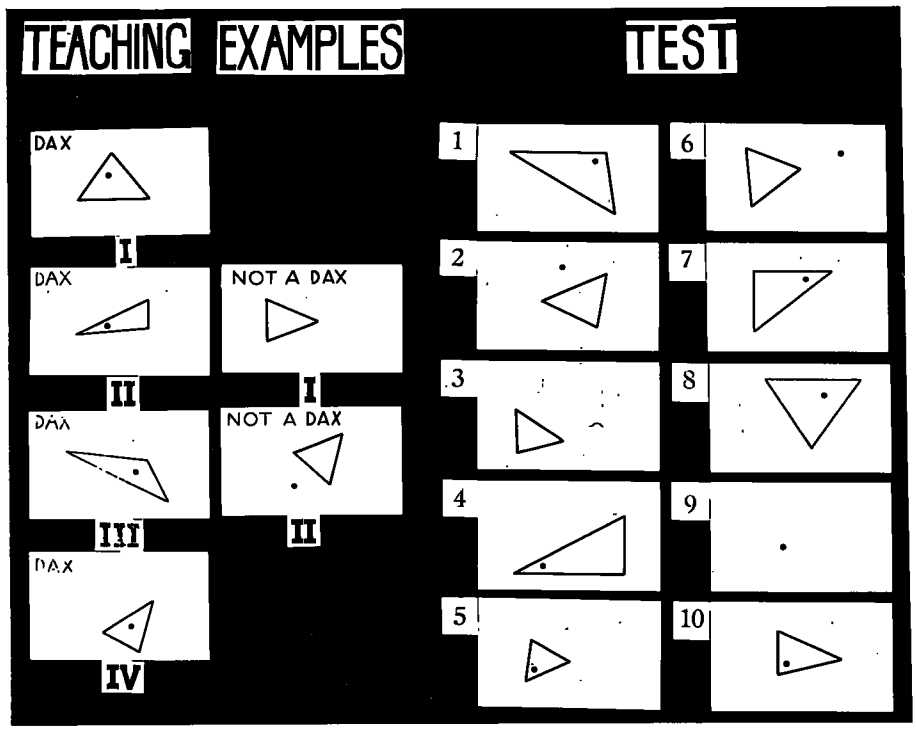
TEACHING EXAMPLES
TEST
DAX
I
DAX
II
NOT A DAX
I
DAX
III
NOT A DAX
II
DAX
IV
1
2
3
4
5
6
7
8
9
10

Fig. 1

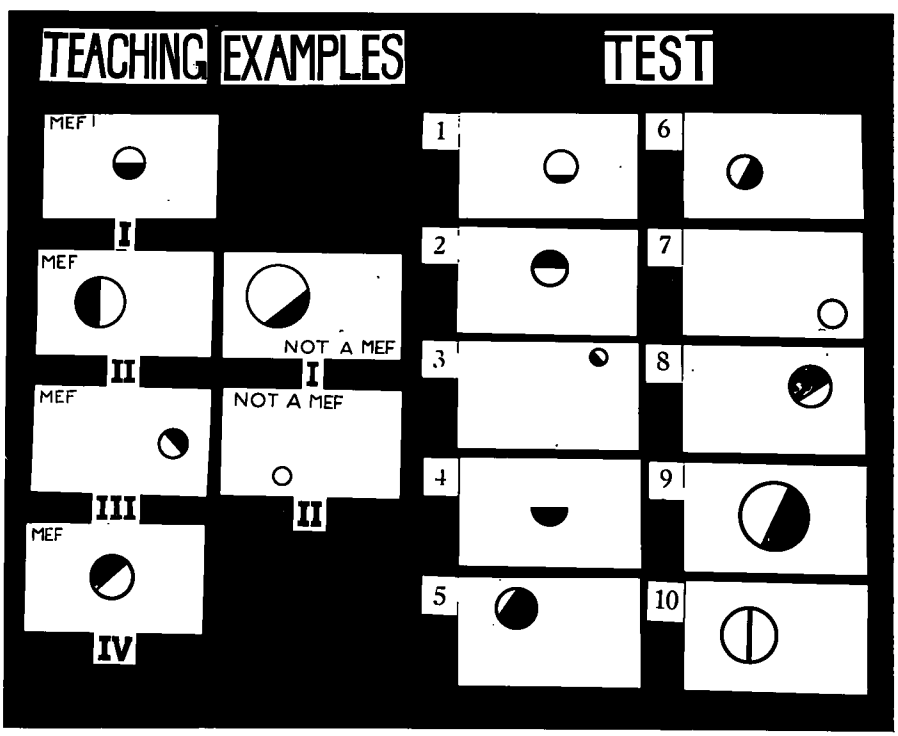
TEACHING EXAMPLES
TEST
MEF
I
MEF
II
NOT A MEF
I
MEF
III
NOT A MEF
II
MEF
IV
1
2
3
4
5
6
7
8
9
10

Fig. 2

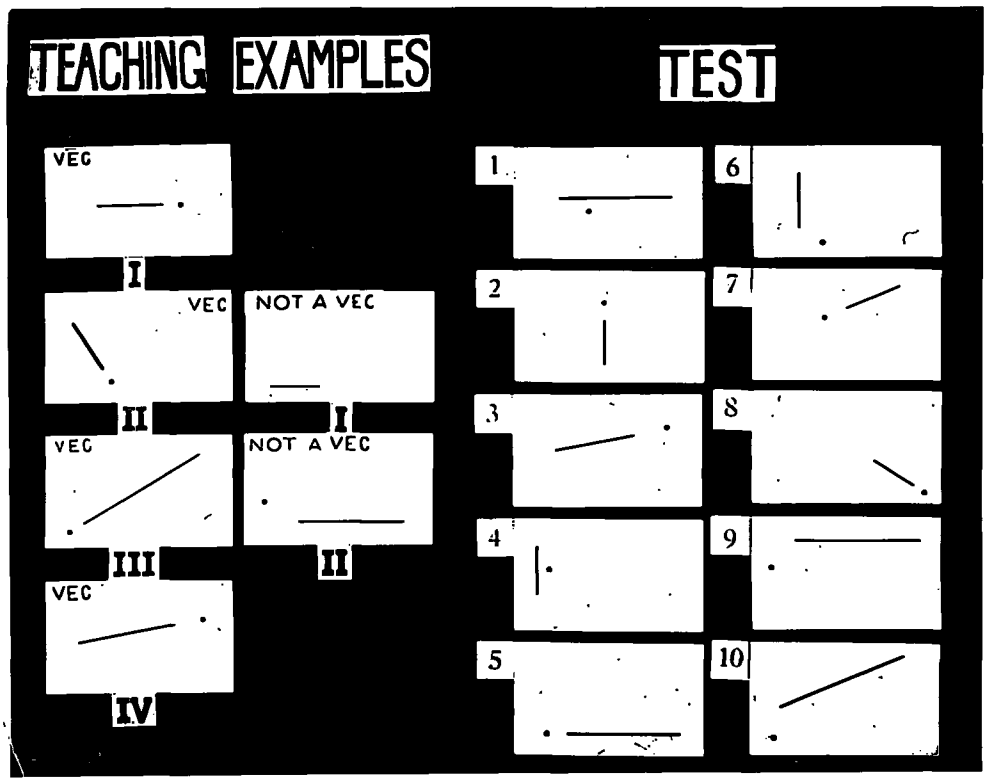
TEACHING EXAMPLES
TEST
VEC
I
VEC
NOT A VEC
II
I
VEC
NOT A VEC
III
II
VEC
IV
1
2
3
4
5
6
7
8
9
10

Fig. 3

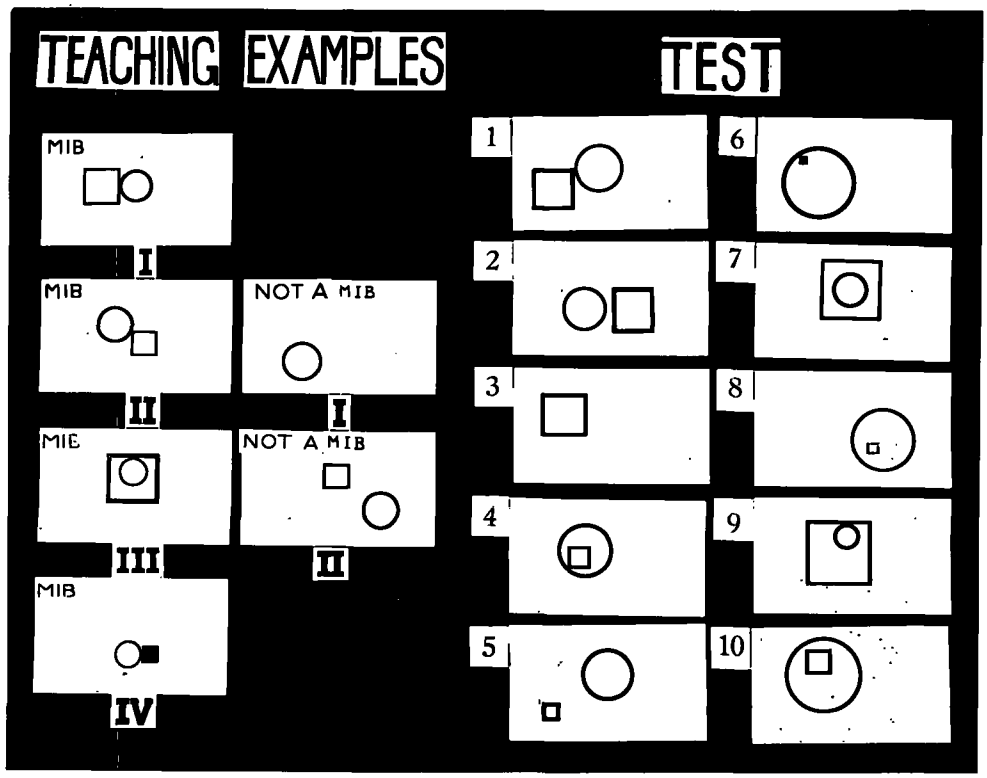
TEACHING EXAMPLES
TEST
MIB
I
MIB
NOT A MIB
II
I
MIB
NOT A MIB
III
II
MIB
IV
1
2
3
4
5
6
7
8
9
10

Fig. 4

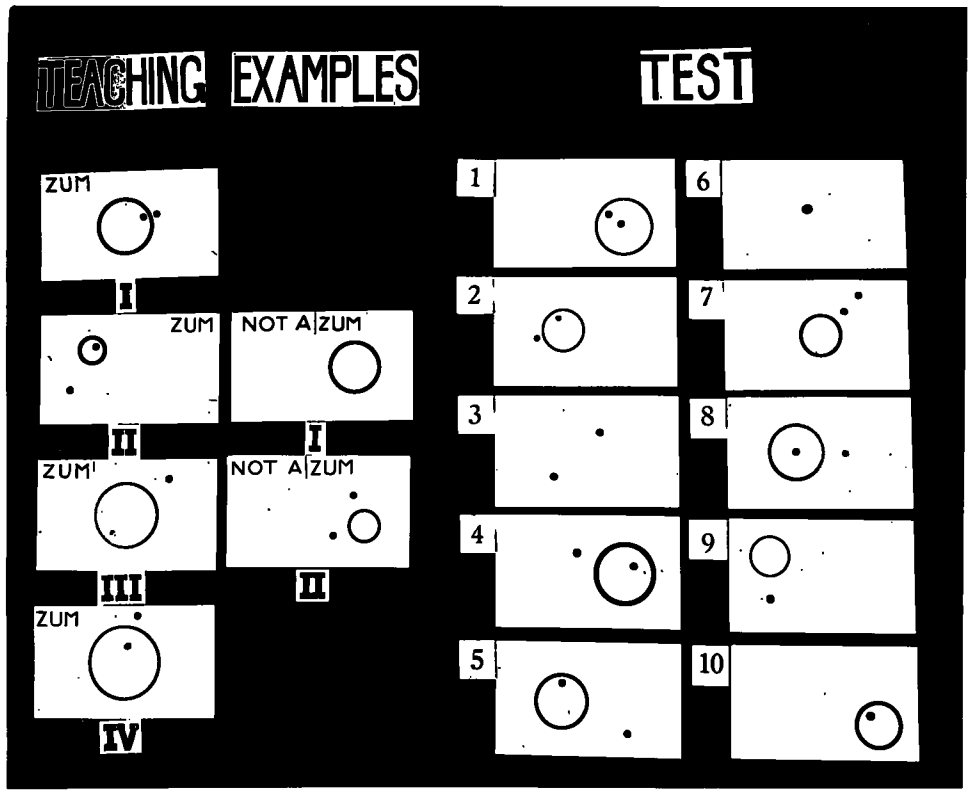
TEACHING EXAMPLES
TEST
ZUM
I
ZUM
II
NOT A ZUM
I
ZUM
III
NOT A ZUM
II
ZUM
IV
1
2
3
4
5
6
7
8
9
10

Fig. 5

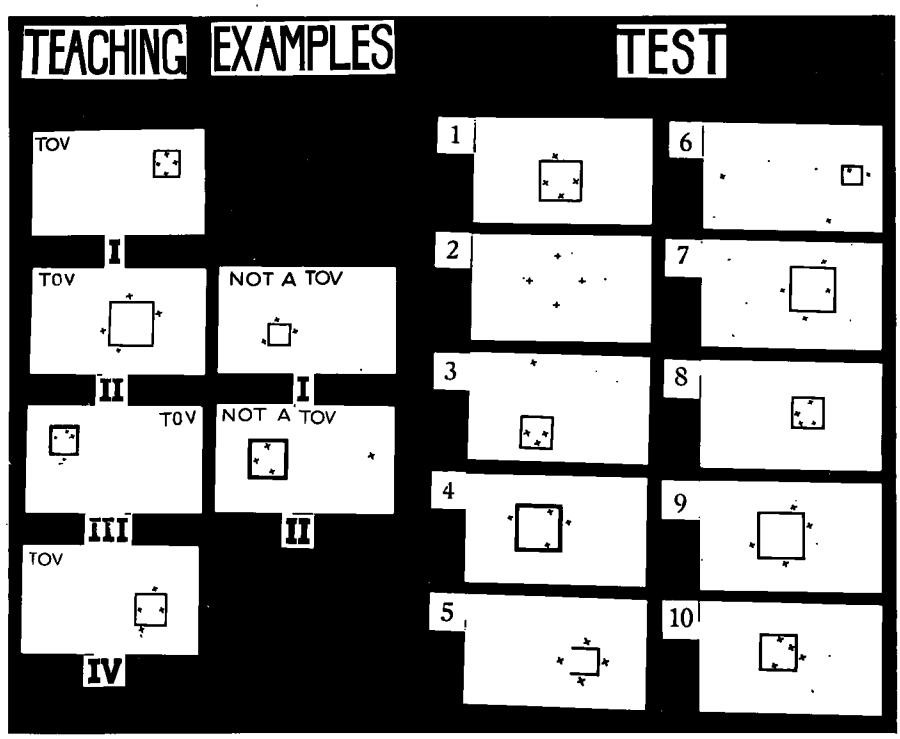
TEACHING EXAMPLES
TEST
TOV
I
TOV
II
NOT A TOV
I
TOV
III
NOT A TOV
II
TOV
IV
1
2
3
4
5
6
7
8
9
10

Fig. 6

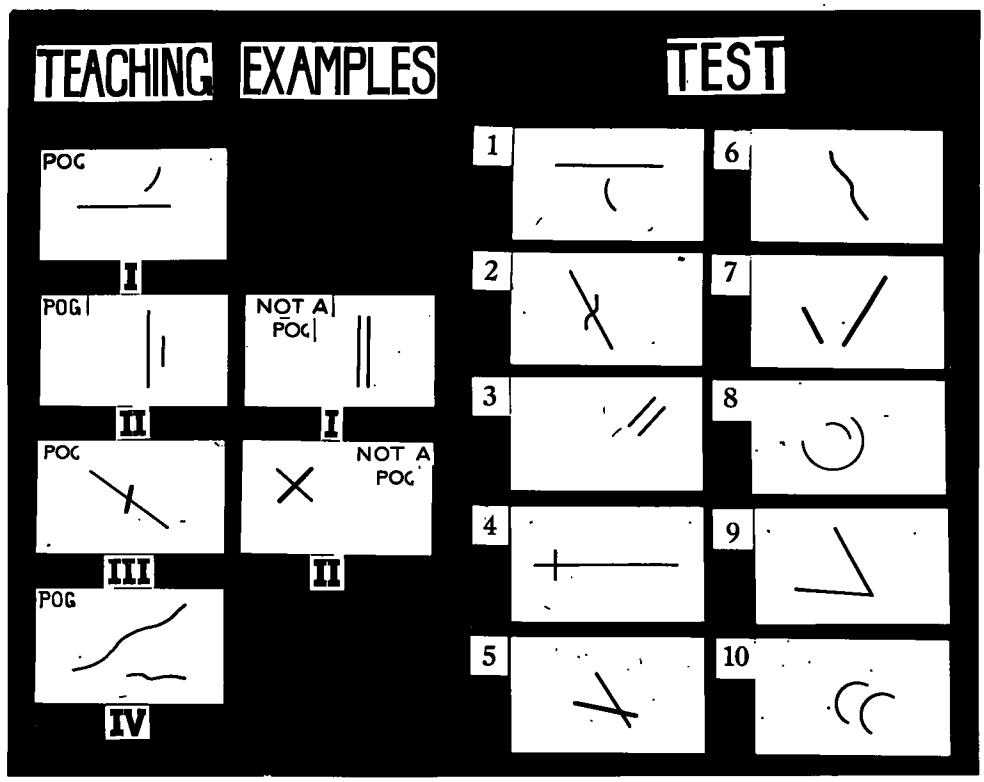
TEACHING EXAMPLES
TEST
POG
I
POG
II
NOT A POG
I
POG
III
NOT A POG
II
POG
IV
1
2
3
4
5
6
7
8
9
10

Fig. 7

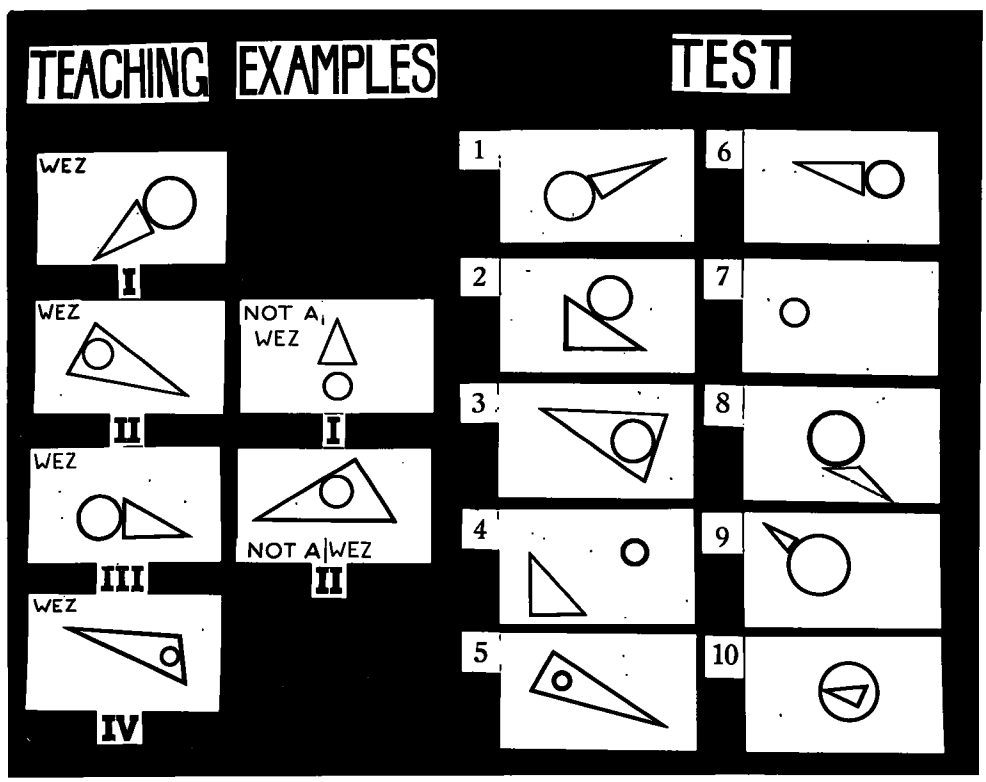
TEACHING EXAMPLES
TEST
WEZ
I
WEZ
II
NOT A WEZ
I
WEZ
III
NOT A WEZ
II
WEZ
IV
1
2
3
4
5
6
7
8
9
10

Fig. 8

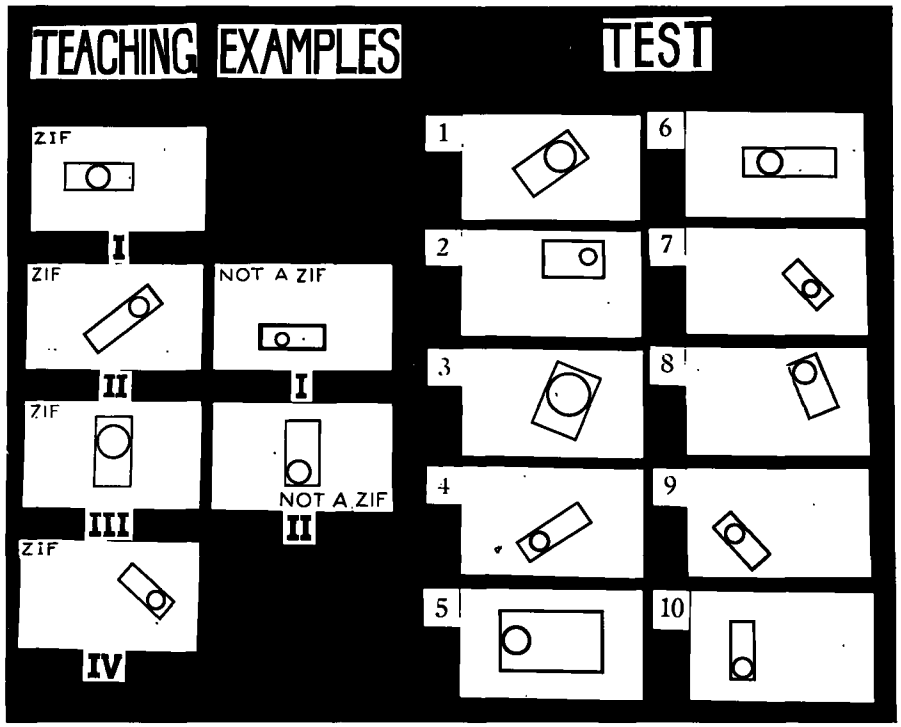
TEACHING EXAMPLES
TEST
ZIF
I
ZIF
NOT A ZIF
II
I
ZIF
NOT A ZIF
III
II
ZIF
IV
1
2
3
4
5
6
7
8
9
10

Fig. 9

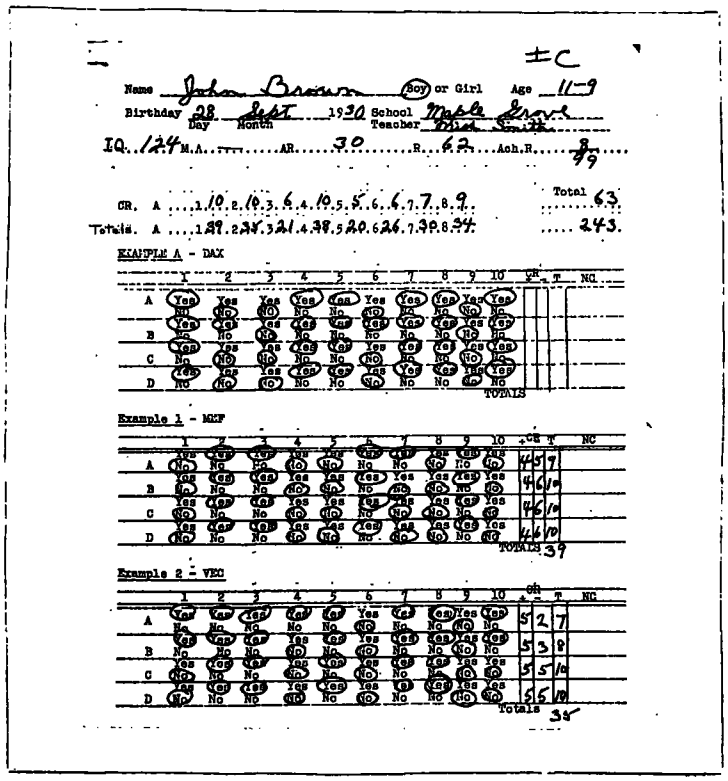
±C
Name John Brown Boy or Girl Age 11-9
Birthday 28 Sept 1930 School Maple Grove
Day Month Teacher
IQ 124 M.A. AR 30 R. 62 Ach.R. 8/9
CR. A 1.10 2.10 3.6 4.10 5.5 6.6 7.7 8.9 Total 63
Totals. A 1.89 2.38 3.21 4.38 5.20 6.26 7.39 8.34 243.
EXAMPLE A - DAX
Example 1 - MEF
Example 2 - VEC
TOTALS

Fig. 10

The teaching and test instances related to each of the nine "concepts" included both positive and negative instances. Positive examples or instances of a given concept referred to those patterns embodying the relationship which defined the concept in question, while negative examples or instances referred to patterns in which this relationship was absent. Positive examples or instances differed from one another in the size and position of the elements, and in heaviness of outline; negative examples or instances, besides being dissimilar in these respects, violated one or more of the conditions demanded by the concept. A clearer conception of the material employed in this study may be had by reference to Figs. 2 - 20. As regards the tests, positive and negative instances were arranged in chance order, one test containing as many as six positive instances, several containing five such instances, while the remainder had but four.

This study was divided into three experiments, each of which may be outlined briefly. In Experiment I four positive teaching examples were serially presented one at a time. Each was exposed for a study-period of 8 seconds, after which it was removed and followed immediately by the test bearing a time-limit of 25 seconds. This manner of introducing and exposing the teaching examples will be designated by the term "successive presentation". In

Experiment II the procedure was identical with that just described except that the four teaching examples, instead of being positive throughout, included an equal number of positive and negative examples. Presented alternately, each positive and negative example was submitted for study, and upon its removal was followed by the test. Experiment III employed the same teaching examples as in Experiment II, but differed from the latter in the method of presentation. Each example was presented together with those which preceded it, and, in addition to the prescribed 8 seconds of exposure, was permitted to remain before the subjects during the whole of the testing period, thereby greatly reducing the effect of memory upon the learning of the concepts. This system of presenting the teaching examples will be referred to as "cumulative presentation".

The three experiments were distinguished from one another, therefore, in respect to the teaching method applied. In point of similarity, however, all experiments employed precisely the same tests, each test being repeatedly presented subsequent to the study of the teaching examples. Also, in accordance with the need to control all factors likely to influence test procedure, a standard set of instructions accompanied each experiment. In each case, the subjects were introduced to the problems by an illustrative example, the concept "Dax", through a series of steps comparable to those to be employed in learning the

concepts comprising the test. Finally, the subjects were warned not to change their answers to an item after the next test item appeared. No further assistance beyond the preliminary instructions was given at any point in the experiments.

In conclusion, the three experiments may be classified as follows:

Experiment I:	A group study of the effect upon concept formation of the successive presentation of positive teaching examples.
Experiment II:	A group study of the effect upon concept formation of the successive presentation of alternate positive and negative teaching examples.
Experiment III:	A group study of the effect upon concept formation of the cumulative presentation of alternate positime and negative teaching examples.

2. Apparatus.

In contrast to the presentation methods used in the previous studies, the stimuli were submitted to groups of subjects by means of lantern slides. Two projectors were employed, one to flash on the teaching examples, the other the test instances. The experiments were conducted in the schools, a classroom or small auditorium being set aside for the purpose and partially darkened, but permitting of sufficient light for the recording of answers. Total time required for administering the tests was approximately 35 minutes.

In making up the film slides the necessary patterns were drawn on plain white cards. Six teaching cards, four positive and two negative, and ten test cards were drawn up for each concept, the total number of cards being 144. These were then numbered and labelled, arranged in the desired order, and photographed. The negative film was used in making up the slides, with the result that all figures were projected upon the screen as white against a black background.

Response was recorded in triple-page booklets, the first page of which is shown in Fig.10. Space was allotted in which the subject was required to fill in his (or her) name, sex, and school, further provision being made for additional data to be inserted by the experimenter. On the first page, as on the two succeeding pages, space was provided for responding to three concepts, each in order of presentation. Under each item number the first, second, third, and fourth presentations of the test were labelled A, B, C, and D, respectively. Henceforth throughout this study it will be found convenient so to designate the several presentations of the test. The numbers 1, 2, 3,.....10 corresponded to the ten instances of the concept, positive and negative, which constituted the test. The subject's task was to determine whether a given test instance was or was not representative of a particular concept, and to draw a circle around either "yes" or "no" accordingly.

This procedure is outlined in detail in Appendix I.

In scoring the results, use was made of the three columns at the right. These columns were labelled "+", "-", and "T", from left to right, and were reserved in that order for scores based upon the number of correct recognitions of positive instances, of negative instances, and of the total of positive and negative instances. In the following chapters, except where specific mention is made of "+" and "-" component scores, discussion of test performance will have reference solely to "T" scores.

3. Subjects

The present study was conducted with the collaboration of the Superintendent and of the principals and teachers of nine Vancouver schools. The tests were administered in June 1942 to Grade VI children of white extraction. Selection of schools was such as to provide a fair distribution of socio-economic factors.

In order to insure a suitable level of difficulty for each test item and to determine the adequacy of the instructions, three trial experiments were conducted with 60 Grade VI subjects. The experimental groups upon which the analysis is based were limited to 270 of a total of 440 subjects originally tested. This reduction arose from the need for making up three comparable groups, one for each of the three experiments. Each group was numbered

according to the experiment in which it participated; thus, those subjects engaged in Experiment I formed Group I, those engaged in Experiment II formed Group II, and so on. These groups were composed by matching individuals on the basis of sex, chronological age, and I.Q. as measured by the Otis Self-Administering Intermediate Examination. Boys and girls involved in one experiment were matched with one another and with boys and girls taking part in each of the two remaining experiments. Since sex was among the factors determining classification, it is obvious that there must be altogether 6 experimental groups, each containing 45 subjects.

The average intelligence of the 6 groups thus formed is shown in Table I. Critical ratios of the differences between means and standard deviations of any two groups did not exceed .38. The average chronological age of each of these groups is listed in Table II. Here again differences were statistically negligible. In matching it was found impracticable to employ a range smaller than ±5 I.Q. points and 5 months chronological age. For example, a boy in Group I who possessed an IQ rating of 112 and a chronological age of 12 years 3 months was matched with subjects in each of the other 5 groups whose IQ's fell within the range 107 - 117, the further requirement being that the difference between the

TABLE I. AVERAGE INTELLIGENCE OF MATCHED GROUPS. STANDARD DEVIATIONS.

	GROUP I		GROUP II		GROUP III	
	Boys	Girls	Boys	Girls	Boys	Girls
A.M...	110.87	110.54	111.08	111.08	110.60	110.54
S.D...	9.78	10.35	9.90	9.60	10.29	9.78

the chronological ages of any two of the subjects thus matched not exceed 5 months. Intelligence ratings were obtained from tests administered to all subjects earlier in the year; chronological ages were listed as of June 30, 1942.

Since the matter of distinguishing between the performances of subjects of high intelligence rating and those of low intelligence rating was of considerable significance for this study, it was also decided to subdivide each of the 6 experimental groups into high, medium, and low IQ groups, as indicated in Table III.

TABLE II. AVERAGE CHRONOLOGICAL AGE OF MATCHED GROUPS. STANDARD DEVIATIONS. (EXPRESSED IN MONTHS)

	GROUP I		GROUP II		GROUP III	
	Boys	Girls	Boys	Girls	Boys	Girls
A.M...	145.56	145.62	145.62	145.34	145.34	145.62
S.D...	5.58	5.44	5.10	5.38	4.62	5.40

TABLE III. AVERAGE INTELLIGENCE OF SUB-GROUPS. STANDARD DEVIATIONS.

		GROUP I		GROUP II		GROUP III	
		Boys	Girls	Boys	Girls	Boys	Girls
HIGH GROUP	A.M.	121.10	121.77	121.63	121.23	121.90	121.50
	S.D.	4.74	4.44	4.64	3.92	4.52	4.00
MED. GROUP	A.M.	111.50	110.83	111.37	111.77	110.57	110.70
	S.D.	3.66	3.90	2.12	3.78	3.92	3.56
LOW GROUP	A.M.	100.17	99.23	100.03	100.57	99.37	99.90
	S.D.	4.42	5.74	4.64	4.66	5.44	4.88

Thus, the resulting groups each involved 15 subjects. The maximum difference in average intelligence between like groups had a critical ratio of .82. Differences in variability of IQ between like groups or between like and unlike groups were somewhat greater, though none was significant.

Average chronological ages pertaining to the groups in question are listed in Table IV. In each instance it will be observed that the greatest differences in average chronological age were to be found between the high and low groups, though again these differences were not statistically significant. Considered horizontally and vertically, differences in variability between groups yielded a maximum critical ratio of 1.23. For present purposes these sub-groups are sufficiently well equated

TABLE IV. AVERAGE CHRONOLOGICAL AGE OF SUB-GROUPS. STANDARD DEVIATIONS. (EXPRESSED IN MONTHS)

		GROUP I		GROUP II		GROUP III	
		Boys	Girls	Boys	Girls	Boys	Girls
HIGH GROUP	A.M.	142.77	144.37	144.10	143.30	143.70	144.23
	S.D.	4.80	4.20	5.00	4.00	3.80	4.60
MED. GROUP	A.M.	145.97	144.63	144.50	144.77	144.63	144.63
	S.D.	6.00	5.80	5.00	5.60	4.60	5.40
LOW GROUP	A.M.	147.97	147.83	148.23	147.97	147.70	147.97
	S.D.	4.60	5.40	4.20	5.40	4.40	5.40

to provide some indication of group performance in relation to intelligence, though the inconstancy of the age factor, together with the small number of subjects involved in each case, render any conclusions based thereon as merely suggestive of certain trends in performance.

CHAPTER III.

THE EXPERIMENTS

1. The D Score As A Basis For Analysis.

In conducting a quantitative analysis of the data, the first question concerns the particular score that is to serve as axbasis for interpretating results. Is it possible to find an approach to the study of test performance which combines maximum validity with minimum computation? For example, is the sum of the scores on the four tests for all eight concepts to provide the basis from which our conclusions derive, or is there some other standard equally acceptable, but which lends itself more readily to calculation? In an effort to provide a satisfactory answer to the problem, it was decided to compute the correlations between the total of the A, B, C, and D scores and the D score*. The results are set forth in Table V. It may be noted that slightly higher correlations were found in the case of Group I in which the negative teaching examples were absent, but in general it appears that this study may well be based upon an analysis of the D score.

TABLE V. CORRELATIONS BETWEEN D SCORE AND TOTAL SCORE. STANDARD ERRORS.

	GROUP I		GROUP II		GROUP III	
	Boys	Girls	Boys	Girls	Boys	Girls
r....	.94	.92	.89	.90	.91	.90
SE_r..	.02	.02	.03	.03	.03	.03

* Here, as later, reference to the A score implies the total of scores on the A test for eight concepts, unless otherwise stated. The same applies in regard to B, C, and D scores.

2. Sex Differences.

With the establishment of the D score as the basis for analysis, the next step calls for a comparison of sex groups to determine the advisability of continuing to treat these as separate units or of combining their results for each experiment. The answer to the question of sex differences is provided by the critical ratios of Table VI in which mean scores and standard deviations are compared.

In both cases involving successive presentation boys showed only a slight tendency to exceed the girls, this tendency being most evident in Group I. On the other hand, in the case of cumulative presentation the girls achieved the highest mean score. Though Table VI makes no mention of the fact, satisfactory significance* (critical ratio of 2.02) characterized the difference between achievement of boys and of girls, to the advantage of the latter.

TABLE VI. CRITICAL RATIOS OF SEX DIFFERENCES IN MEAN SCORES AND VARIABILITY ON THE D TEST.

	GROUP I		GROUP II		GROUP III	
	Boys	Girls	Boys	Girls	Boys	Girls
A.M....	53.57	52.86	48.50	47.52	51.43	57.12
$C.R._M$...	.80		.35		2.21	
A.D....	7.08	8.40	13.48	13.24	14.40	9.48
$C.R._\sigma$...	1.14		.12		2.70	

*Henceforth, critical ratios of 1.65, 2.35, 3.00 will constitute the lower arbitrary limits for satisfactory, high, and virtual statistical significance, respectively. See Peters and Van Voorhis, Statistical Procedures and Their Mathematical Bases, pp. 138, 176.

Thus,an overall comparison of boys with girls indicates that while the girls made the lowest average scores, they also attained the highest average on the D Test. However, while the results offer no conclusive evidence of marked sex differences in the handling of concepts, the extent of the differences in mean scores and variability between boys and girls in Group III justifies treating the sexes separately throughout the remainder of this study.

3. Group Differences.

Group differences are next examined to determine the effect of variations in the method of instruction. The necessary data for this purpose are furnished by Table VII. The groups whose differences are under study in each case are indicated in the column at the extreme left, the remaining

TABLE VII. GROUP DIFFERENCES IN MEAN SCORES AND VARIABILITY ON THE D TEST. CRITICAL RATIOS.

GROUPS	BOYS				GIRLS			
	Mean		S.D.		Mean		S.D.	
	DIFF.	C.R.	DIFF.	C.R.	DIFF.	C.R.	DIFF.	C.R.
I - II	5.07	+2.23	6.40	-3.98	5.34	+2.28	4.84	-2.92
I - III	2.14	+.89	7.32	-4.33	4.26	-2.27	1.08	- .81
II - III	2.93	-1.00	.92	- .44	9.60	-3.95	3.76	+2.19

columns containing the actual differences in mean scores and variability, together with the critical ratios of these differences. A positive critical ratio indicates that the first-named group in the extreme left-hand column attained the higher mean score or greater variability, as the case may be; on the other hand, a negative ratio points to the first-named group as possessing the lower mean score or as being the less variable of the two.

A comparison of the mean scores of all boys' groups revealed Group I as the most successful, Group II as the least successful of the three groups. Difference in mean scores between Groups I and II approached high statistical significance, while that between Groups I and III was considerably less.

Of equal interest is the manner in which the scores were distributed about the mean in the above groups. Differences between standard deviations showed that boys in Group III were scarcely more variable in performance than those in Group II; on the other hand, boys in Groups II and III showed promise of always displaying greater variability on the tests than boys in Group I.

The foregoing results indicate, in the case of the boys, a tendency toward higher mean scores and greater uniformity of response from successive presentation involving only positive examples than from either of the two remaining methods.

Turning now to the girls, those in Group III achieved the highest mean score on the D Test, those in Group II the lowest. The difference between scores in Groups I and scores in Groups II and III approached high significance, while the difference between scores in Groups II and III was virtually significant.

As regards variability, Group II girls showed themselves more variable than Group III girls and decidedly more so than Group I girls, a satisfactory and a high significance attaching to the respective differences. There was little difference in variability between Groups I and III.

Considering only mean scores and disregarding differences in variability, the results suggest the advantage to the girls of the method employing cumulative presentation of both positive and negative examples in the teaching series. The boys, on the other hand, seemed to derive most benefit from successive presentation in which negative examples of the concept were excluded. For both boys and girls successive presentation utilizing the negative example appeared as the least favorable mode of instruction, and in both cases involving positive and negative examples the method of cumulative presentation held the advantage.

While the recommendation of any particular method of presentation would be rather presumptuous at this stage in the analysis, at least one or two facts are worth noting:

Group instruction as herein provided leads to a lower average score and to a greater spread in achievement when assisted by negative examples than when only positive examples are presented. These findings are in sharp contrast to those of Wood in which the presence of negative examples within the teaching series boosted performance and produced a closer grouping of the individual scores about the mean. Dickinson's results yielded no clear-cut tendencies in either direction, though they offered some evidence of a decrease in variability accompanying presentation of the negative example. Dickinson's and the present study advance contradictory claims regarding the effect upon performance of varying only the memory factor. Thus, in the former higher achievement accompanied successive presentation, while in the latter a reverse trend favored cumulative presentation where the negative example was concerned.

4. High And Low IQ Groups Compared.

In concluding this phase of our study, an attempt should be made to determine the relationship between intelligence and concept formation. An insight into relative performance by subjects differing widely in intelligence may be gained by reference to the sub-groups mentioned in the last chapter. We shall find it convenient at this time to limit ourselves to a study of high and low IQ groups, utilizing the combined results of the A, B, C and D tests for each group. Since interest again lies with obtained

TABLE VIIIa. DIFFERENCES IN MEAN SCORES AND VARIABILITY BETWEEN HIGH IQ GROUPS, BASED UPON THE COMBINED TOTAL OF A, B, C, AND D SCORES. CRITICAL RATIOS.

GROUPS	BOYS				GIRLS			
	Mean		S.D.		Mean		S.D.	
	DIFF.	C.R.	DIFF.	C.R.	DIFF.	C.R.	DIFF.	C.R.
I - II	1.33	-.12	8.10	-1.07	4.00	+.34	7.60	-.90
I - III	7.33	-.58	17.10	-1.90	16.67	-1.98	12.00	+2.02
II - III	6.00	-.43	9.00	- .92	20.67	-2.02	19.60	+2.71

TABLE VIIIb. DIFFERENCES IN MEAN SCORES AND VARIABILITY BETWEEN LOW IQ GROUPS, BASED UPON THE COMBINED TOTAL OF A, B, C, AND D SCORES. CRITICAL RATIOS.

GROUPS	BOYS				GIRLS			
	Mean		S.D.		Mean		S.D.	
	DIFF.	C.R.	DIFF.	C.R.	DIFF.	C.R.	DIFF.	C.R.
I - II	28.67	+2.78	17.50	-2.40	16.67	2.08	.70	-.12
I - III	29.33	+2.78	18.50	-2.48	22.66	-2.38	8.10	-1.20
II - III	.66	+.05	1.00	-.11	39.33	-4.08	7.40	-1.09

differences between scores rather than with the actual scores themselves*, only these differences and their critical ratios are tabulated above. As with Table VII, no provision is made for a direct comparison of boys with girls.

Excluding for a moment all comparisons involving Group III girls, it will appear that score differences between high IQ groups were negligible, while those between low groups were highly in favor of Group I. In general, an

* Mean scores and standard deviations for high and low IQ groups are provided in Table A-B, Appendix II.

increase in variability accompanied the introduction of the negative example. These indications suggest that, while the negative example has little effect upon the group performance of bright children, it may actually prove detrimental to those of lesser intelligence under conditions similar to those which prevailed in these experiments.

Table IX presents these differences from another angle by directly comparing the mean performance of high and low groups, with the following results: Scores in Low Groups II were significantly lower than scores in High Groups I, while the differences between scores in Low Groups I and High Groups II were of only satisfactory or negligible significance. All this suggests that the negative example serves merely to re-emphasize the difference in intelligence between high and low groups; that is, the lower the average intelligence of the group, the more inhibitive may become the effect of the negative example upon test performance. In other words, the evidence offers nothing to substantiate

TABLE IX. DIFFERENCES IN MEAN SCORES BETWEEN HIGH AND LOW IQ GROUPS, BASED UPON THE COMBINED TOTAL OF A, B, C, AND D SCORES. CRITICAL RATIOS.

GROUPS	BOYS		GIRLS	
	DIFF.	C.R.	DIFF.	C.R.
I(H) - II(L)	38.67	3.44	43.33	4.65
I(L) - II(H)	11.33	-1.56	22.66	-2.08
II(H) - III(L)	40.66	3.18	0.0	0.0
II(L) - III(H)	46.00	-3.21	60.00	-8.49

Wood's claims for the instructional advantage of the negative example to those of lesser intelligence.

A similar comparison of high and low IQ groups in Groups II and III suggests that azreduction in the memory factor had a negligible effect upon the relative performance of high and low boys' groups; the performance of high and low girls' groups within Group III, however, was far superior to that of other sub-groups on the same intelligence level.

The singular performance of Group III girls as a whole lends a certain inconsistency to the general pattern which is unexplainable in terms of intelligence, arithmetic reasoning, or reading, insofar as can be determined. The possibility that behavior was actuated by certain motivational factors peculiar to one experimental setting is minimized by the fact that the subjects comprising this particular group represented three different schools. It may be that this was a select group in terms of an ability or abilities ignored by previous measurement.

5. Summary of Chapter III.

6 Following is a condensation and restatement of findings up to this point.

1. Average performance on the D Tests correlated highly with average performance on the combined A, B, C, and D tests.

2. Instructions attended by successive presentation performed their function more satisfactorily when negative teaching examples were excluded. For the presentation of both positive and negative examples the cumulative method was the more effective.
3. In general, dispersion or scatter of scores was augmented by the presence of the negative example in the teaching series.
4. Boys showed a tendency to benefit most from instruction by successive presentation of positive examples, girls from instruction by cumulative presentation of positive and negative examples. Successive presentation favoured the boys and cumulative presentation the girls, though no decided sex differences were manifested.
5. Results suggest that group instruction utilizing the negative example had little effect upon the response of bright children, while adversely affecting that of normal children.
6. The conflicting evidence of Wood's findings and of those of the present study points to possible inherent differences which distinguished performance in each of the two settings, and suggests that care must be exercised in attempting to generalize from one to the other.

CHAPTER IV.

TEST RELIABILITY AND VALIDITY

1. Test Reliability.

Of paramount importance in test evaluation is the degree to which consistency of performance characterizes the two halves of a test or is maintained through several presentations of the test or its equivalent. The split-half technique being the only means available for an estimate of reliability in this case, the test was divided into two equal parts, each containing four items corresponding in difficulty to the four items in the other half, in accordance with the underlying assumptions governing this method. To this end it became necessary to reassemble the test items for each of the whole groups I, II, and III, though the same item-arrangement held for boys' and girls' groups within each. Reliabilities were obtained by correlating the two halves so formed, and then applying the Spearman-Brown formula. The resulting values are tabulated below.

TABLE X. CORRELATIONS BETWEEN TEST-HALVES, THEIR STANDARD ERRORS, AND RELIABILITY COEFFICIENTS FOR THE D TEST.

	GROUP I		GROUP II		GROUP III	
	Boys	Girls	Boys	Girls	Boys	Girls
$r_{\frac{1}{2}\frac{1}{2}}$	.62	.69	.79	.71	.84	.68
$SE_{r_{\frac{1}{2}\frac{1}{2}}}$	.09	.08	.06	.07	.04	.08
r_{11}	.77	.82	.88	.83	.91	.81

The fact that these values may be higher than might obtain from the use of equivalent forms constitutes no criticism of the Spearman-Brown formula, according to Jackson and Ferguson[1], but is simply attributable to "the process of splitting the test." In any case, insofar as an estimate is possible, indications point to a fairly high degree of reliability for the type of test under consideration. It is noteworthy that maximum reliability pertained to the two boys' groups subjected to the negative teaching example.

In line with the assumption that the greater the element of chance, the lower the reliability, Symonds[2] contends that response on the basis of the No-Yes choice tends to reduce test reliability. Since reliability is largely a function of variability, the effect upon the former of added opportunity for guesswork is obvious. But in spite of this claim, the relatively high value of the coefficients obtained is justification for concluding that the role of chance has received no undue emphasis in the present tests.

A possible explanation for present reliabilities being slightly lower than those of Dickinson's tests[3] may focus upon the distractive influences resulting from exposure of the two projectors during the testing period, and from the frequent need for conducting the experiments

---

1. Jackson, R.B., and Ferguson, G.A., Studies on the Reliability Of Tests, Bulletin No.12 of the Department of Educational Research, University of Toronto, 1941, p.11.
2. Symonds, P.M., "Factors Influencing Test Reliability", Journal of Educational Psychology, vol. XIX, 1928, p. 79.
3. Dickinson, A.E., op. cit., p. 47.

outside the familiar classroom surroundings. Furthermore, had present reliabilities been based upon correlations between totals of A, B, C, and D scores for each test-half instead of upon correlations between D scores, the values would probably have been somewhat higher.

2. An Aspect of Test Validity: Correlations With Intelligence And Other Variables.

Closely allied to test reliability is the matter of validity. With what success do the present tests accomplish the segregation and measurement of generalizing ability? While the answer to any such question relating to mental tests is necessarily but an estimate of the facts, several means exist for deriving conclusions. These consist in computing correlations betweem the test concerned and some criterion, in studying the validity and intercorrelations of the test items themselves, in applying the index of reliability, or in using any of the other direct or indirect methods for estimating test validity. The first-named, which has found wide application, was the one employed in this study, supplemented later (Chap. V) by an investigation of item validity.

The availability of performance ratings on standard tests of intelligence, reading, and arithmetic reasoning made it desirable to compute correlations between each of these and the tests of generalizing ability to determine whether the latter are measuring abilities covered by the

other tests or whether they are measuring something quite different. At this point a brief description of the arithmetic reasoning test is in order. In this test*, all items entail reading and memory, and a clear demand is placed upon relational and numerical ability. Typical of problems on the arithmetic reasoning test are the following:

(a) Alice has filled 48 pages of her 64-page exercise book. How many pages of her exercise book are still blank?

(b) The discount at 5% on a bill was $20.00. How much was the bill before it was discounted?

(c) A box which has a volume of 24 cubic feet is 4 feet long, 3 feet wide. How deep is it?

(d) A newsboy made $1\frac{1}{2}$ cents on each paper he sold. This was 60% of the cost. What was the selling price of each paper?

(e) A man rows down stream 6 miles in 2 hours and, returning against the current, takes 6 hours. Find his rate of rowing and the rate at which the stream flows.

In each of these problems the role of memory is seen in the recall of certain fundamental rules related to areas, volumes, percentages, subtraction, multiplication, division, and so on. Accuracy in dealing with numbers is also a factor in reaching a solution. Possession of these two factors, memory and accuracy, seems sufficient to produce the desired result at the Grade VI level in the case of easier problems, such as (a), (b) and (c), which merely require a mechanical application of some simple arithmetic rule. Where more difficult problems are concerned, of which problem (e) is an

---

* Vancouver Tests: Reasoning in Arithmetic, Form A.

example. There must be added some relational or integrative process tentatively scknowledged as arithmetic reasoning. It appears, therefore, that somexof the items on the arithmetic reasoning test draw upon reasoning ability.

Correlations between each of these three tests and those of generalizing ability are assembled in Table XK, together with correlations between intelligence and each of reading and arithmetic reasoning. Considering the degree of error involved, the nature of the test material, and the

TABLE XI. ZERO-ORDER CORRELATIONS OF D TEST WITH INTELLIGENCE, READING, AND ARITHMETIC REASONING, AND OF INTELLIGENCE WITH READING AND ARITHMETIC REASONING. STANDARD ERRORS.

	GROUP I				GROUP II				GROUP III			
	Boys		Girls		Boys		Girls		Boys		Girls	
	r	SE_r	r	SE_r	r	SE_r	r	SE_r	r	SE_r	r	SE_r
D & I.Q.	.30	.14	.45	.12	.36	.13	.37	.13	.41	.12	.37	.13
D & R.	.26	.14	.34	.13	.23	.14	.26	.14	.44	.12	.30	.14
D & A.R.	.32	.13	.51	.11	.27	.14	.19	.14	.25	.14	.13	.14
I.Q. & R.	.66	.08	.66	.08	.64	.09	.70	.08	.71	.07	.74	.07
I.Q.&A.R.	.30	.14	.69	.08	.64	.09	.56	.10	.44	.12	.55	.10

size and selectivity of the groups, the results suggest a relationship between D Test scores and performance on the Otis Test. The correlations of the D Test with each of the remaining variables were somewhat lower, for the most part. Intelligence seemed most closely associated with reading ability and least with generalizing ability. By correcting

TABLE XII. CORRELATIONS OF TABLE IX CORRECTED FOR ATTENUATION*.

	GROUP I		GROUP II		GROUP III	
	Boys	Girls	Boys	Girls	Boys	Girls
D & I.Q.	.35	.51	.40	.42	.45	.43
D & R	.31	.40	.26	.30	.49	.35
D & A.R.	.39	.59	.30	.22	.28	.15

for attenuation and so cancelling the distortive effect of chance errors in correlated tests, the coefficients appear as in Table XII.

Results show that D Test performance, under the influence of changes in methods of instruction, exhibited a practically constant relationship with intelligence. This fact becomes even more apparent upon combining and averaging values for boys and girls groups within each of the three major groups. All in all, interrelations with reading ability and arithmetic reasoning may be similarly described, although there is some indication that the correlations of D scores with arithmetic reasoning were lower in Groups II and III than in Group I. If, therefore, the arithmetic reasoning test be accepted as an adequate means for measuring generalizing ability, it appears that the introduction of the negative example impairs the validity of the D Test as a measure of this ability.

From Table XI it would appear that, for Grade VI children, the reading factor in the Otis Test is an important determinant behind intelligence ranking.

* Reliability of Reading and Arithmetic Reasoning Tests was .90.

Reading may therefore supply the reason for the low intercorrelations of D Test scores with intelligence, for language forms an integral part of each of the 75 items on the Otis Test but is confined to the preliminary instructions in the tests of generalizing ability. This explanation applies also to the low relationship between the D and reading tests. On the other hand, the fact that the correlations between these two tests were positive might indicate the presence of a common reasoning factor in each. Or equally probable, the existence of a verbal factor common to both tests may account for the positive correlations, particularly since comprehension of the verbal instructions at the outset was prerequisite to a successful manipulation of the concepts in the generalizing tests.

A full treatment of this aspect of test validity should explore the possibility of a relationship with estimated classroom performance. D Test scores were graded according to the system used in rating school achievement. Those subjects among the best five percent received a grade of A, the next ten percent a grade of B, and so forth. By applying the chi-square test, positive evidence of a relationship between school achievement and D Test performance was established and revealed to be largely, though not entirely, independent of chance factors. These data, together with their expression in terms of the

contingency coefficient (C), are contained in Table XIII. Quite a high degree of association is indicated by the coefficients, but these values must be accepted with certain reservations. Firstly, school achievement rating, instead of being wholly objective in nature, is in part the product of personal judgment. And secondly, since an A grade at one school might carry only B credit at another, and since each experimental group included subjects drawn from a number of schools, it can afford but a rough measure of a subject's standing within that group. Therefore, while the facts support the probability of a positive relationship, its actual extent is problematical. Quite apart from other considerations, a set of low correlations would not have been surprising in view of the large number of abilities governing school work.

TABLE XIII. THE RELATIONSHIP OF D TEST PERFORMANCE AND SCHOOL ACHIEVEMENT IN TERMS OF PROBABILITY (P), CONTINGENCY COEFFICIENTS (C) AND THEIR STANDARD ERRORS (APPROXIMATE).

	GROUP I		GROUP II		GROUP III	
	Boys	Girls	Boys	Girls	Boys	Girls
P.....	.1190	.2611	.1685	.3594	.3256	.1314
C.....	.61	.69	.62	.65	.64	.71
SE_C....	.15	.15	.15	.15	.15	.15

A comparative study of performance on the D Test and on various criteria has thus demonstrated that the relationship between scores on the first-named and those on each of the other tests exhibits a certain semblance of consistency, and in so doing yields some proof of the validity of the tests of generalizing ability. While the results for validity are not entirely conclusive, particularly as regards the form in which the tests were administered to Groups II and III, much of the evidence implies that the present tests were measuring certain qualities beyond the range of the other tests considered.

3. Summary of Chapter IV.

1. Indications suggest that measurement of group performance by these tests is attended by a fairly high degree of reliability.
2. In general, D Test performance exhibited a positive, though not significant relationship with intelligence (as measured by the Otis Test). On the other hand, intelligence seemed to have more in common with reading ability and arithmetic reasoning than with performance on the D Tests.
3. Correlations of D Test performance with reading ability and arithmetic reasoning were generally positive but low.

4. D. Test performance displayed most in common with scholastic achievement, although no accurate measurement of this relationship was possible.

CHAPTER V.

ITEM VALIDITY AND ANALYSIS

Transferring from validation techniques of the type used in the foregoing chapter to an application of "indirect" methods which restrict analysis to details within the test, the next point of consideration is that of item validity, for a test is no morevvalid than the items which comprise it. In this study test items must not be confused with test instances; the term "test item" is herein used to designate the whole battery of test instances, positive and negative of a given concept.

1. Item-Difficulty.

Thurstone[1], summarizing the results of several experiments with Grade VI children, claims that tests emboyding items with a difficulty range extending from approximately 30 percent to 70 percent successes and averaging about 50 percent successes probably carry a higher validity value than tests whose ranges of item-difficulty vary from 80 to 100 percent successes. To obtain an over-all picture of the difficulty-order held by items in our study, rank order of difficulty was determined for the total of A, B, C, and D scores within each item rather than for the D score alone. This step was deemed desirable,

1. Thurstone, T.G., "The Difficulty of a Test and Its Diagnostic Value", Journal of Educational Psychology, vol. XXXII, 1932, pp.341-2.

TABLE XIV. ORDER OF ITEM-DIFFICULTY IN TERMS OF PERCENTAGES OF MAXIMUM POSSIBLE SCORES AND THE NUMBER OF FAILURES (F)*.

GROUP I						GROUP II						GROUP III					
Boys			Girls			Boys			Girls			Boys			Girls		
Item	%	F	Item	%	F	Item	%	F	Item	%	F	Item	%	F	Item	%	F
Zum	84	0	Zum	82	0	Zum	74	4	Zum	76	2	Zum	77	4	Zum	88	0
Vec	81	0	Vec	73	2	Vec	68	10	Vec	66	12	Vec	69	8	Vec	80	1
Mef	68	4	Mef	71	3	Zif	63	11	Wez	64	2	Wez	68	3	Zif	73	6
Zif	63	7	Mib	62	5	Wez	62	5	Zif	61	12	Mef	67	8	Mef	72	2
Wez	62	2	Zif	61	13	Mef	61	16	Mef	59	14	Zif	66	16	Wez	70	3
Mib	59	7	Wez	60	3	Mib	56	11	Mib	58	13	Pog	58	7	Pog	62	3
Tov	57	6	Tov	60	4	Tov	53	14	Pog	56	6	Mib	56	8	Mib	60	7
Pog	56	5	Pog	55	4	Pog	53	11	Tov	54	7	Tov	52	15	Tov	56	6

Failure: a score below 50 percent of the possible score

since the interrelation of D and total scores was known only for the test as a whole, and not for the individual items.

Table XIV lists these items in order of difficulty, least to greatest, together with the percentage of maximum possible score and the number of subjects receiving less than a 50 percent score on each. Errors ranged from 12 to 48 percent of the possible score, depending upon the item and the conditions under which it was presented. Varying the method of instruction did not materially change the rank-order; Zum and Vec retained their positions throughout as the easiest items, while Mib, Pog, and Tov for the most part were the hardest of the series. There was no evidence to show that a given item was learned more effectively by one method than by another, although in no case did the highest average score on an item occur under positive-negative successive presentation.

Analysis of similar data for sub-groups (see Appendix II, Tables C and D) reveals little beyond the fact that there occurred among the high groups a greater spread between average scores on the easiest and most difficult items. Then, too, the closer similarity between orders of item-difficulty among high groups is suggestive of the more predictable manner in which members of these groups may have attacked the problems.

2. Validity And the Diagnostic Value of An Item.

In order to determine the validity of items composing a test, Kelley[1] recommends selecting two groups made up of the 27 percent of the subjects who received the highest scores on the test and the 27 percent who received the lowest scores. For purposes of the present study upper and lower groups were selected from the boys in Groups I and III on the basis of combined A, B, C, and D scores.

The results of this analysis (Table XV) suggest in general greater discriminatory properties for those items listed in Table XIV which are more remote from the 50 percent difficulty level than for items such as Mib, Tov, and Pog which closely approach this level. Rank order correlations between order of difficulty and diagnostic value were .74 and .79 for Groups I and III, respectively, indicating that for the difficulty of items in this study, the easier the item, the greater is likely to be its discriminatory value.

It may therefore be that the optimum difficulty-level approximates 75 percent successes for the material of these experiments. This does not necessarily imply a

---

1. Kelley, T.L., cited in Long, J.A. and Sandiford, P., "The Validation of Test Items", Bulletin No.3 of the Department of Educational Research, University of Toronto, 1935, p. 94.

TABLE XV. SUPERIORITY OF UPPER OVER LOWER GROUPS IN PERFORMANCE ON TEST ITEMS, EXPRESSED IN TERMS OF CRITICAL RATIOS OF THE DIFFERENCES IN MEAN SCORE AND IN VARIABILITY.

	GROUP I BOYS		GROUP III BOYS	
	$C.R._M$	$C.R._\sigma$	$C.R._M$	$C.R._\sigma$
MEF.....	4.30	.47	6.77	.39
VEC.....	8.85	2.62	5.86	1.17
MIB.....	1.00	1.95	3.10	.98
ZUM.....	5.48	.78	7.88	2.65
TOV.....	1.21	.09	2.47	2.21
POG.....	1.40	2.24	4.65	1.41
WEZ.....	5.20	2.87	7.05	1.15
ZIF.....	6.67	1.98	7.83	1.52

contradiction of Thurstone's results, for the present tests are not unlike achievement tests of the true-false type in which the medium difficulty level approaches 75 percent of the possible score. Nevertheless, it would be fallacious to presume the general application of present findings without calling attention to their limitations as defined by the size and selectivity of the groups involved. For, as Long and Sandiford caution, "...the validity values obtained from data gathered on a particular group of subjects are not highly reliable indications of their validities for another and widely different group."[1] In consequence, evidence in Table XV of the greater validity of items listed under Group III may be more apparent than real.

1. Long, J.A. and Sandiford, P., op. cit., p. 107.

TABLE XVI. PERCENTAGE ACCURACY OF POSITIVE AND NEGATIVE COMPONENT SCORES ON THE TOTAL OF TESTS A, B, C, AND D FOR SIX TEST ITEMS.

	GROUP I				GROUP II				GROUP III			
	Boys		Girls		Boys		Girls		Boys		Girls	
	+	-	+	-	+	-	+	-	+	-	+	-
Mef...	73	64	68	75	57	63	54	62	62	70	64	77
Vec...	75	88	60	86	57	78	60	73	66	72	76	82
Zum...	72	92	60	95	62	83	63	83	68	84	79	93
Tov...	33	76	34	78	21	74	21	76	21	73	27	74
Pog...	28	82	25	85	26	80	26	86	31	85	32	94
Zif...	71	49	57	67	63	63	60	62	69	62	74	70

3. Items Analyzed.

Associated with this whole conception of test validity is the need for an investigation of the various factors which combine to make an item easy or difficult. Examination of Table XVI, which translates averaged "no" and "yes" scores for six of the eight test items into percentages based upon accuracy of response to positive and negative test instances, reveals a preponderant tendency to respond more accurately to negative than to positive test instances. This behavior characterized all groups and was especially magnified in the case of the more difficult items. In Mef and Zif only was there any evidence of an equal or greater percentage of "yes" scores; the reason for this lies not so much in the fact that the positive instances of Mef and Zif were more easily identified than were those of other concepts but rather in the fact that relatively fewer negative instances were recognized. In all other items, however, positive instances offered greater difficulty.

Group reaction to specific positive instances in a number of items merits some attention at this point. Since space does not permit a complete study of the material at hand, consideration will be restricted to several outstanding features of the more difficult items as applied to boys in all three groups.

TABLE XVII. NUMBER OF CORRECT RESPONSES TO INDIVIDUAL POSITIVE INSTANCES OF POG, MIB, AND WEZ ON EACH OF TESTS A, B, C AND D. (BOYS' GROUPS ONLY)

		GROUP I					GROUP II					GROUP III				
Instance:		1	2	4	7	8	1	2	4	7	8	1	2	4	7	8
POG	A..	32	2	2	4	4	30	3	2	3	2	41	2	2	0	3
	B..	19	4	9	24	4	22	9	6	12	7	31	7	4	7	4
	C..	17	17	32	12	3	21	5	7	20	6	32	5	17	27	8
	D..	19	15	15	12	12	23	7	12	21	8	32	8	12	21	10
Instance:		1	4	6	9		1	4	6	9		1	4	6	9	
MIB	A..	39	6	1	3		23	4	1	4		20	3	2	5	
	B..	30	6	3	6		23	12	9	7		19	7	10	10	
	C..	19	16	2	19		31	8	1	6		40	9	2	5	
	D..	22	16	1	13		25	10	5	9		35	11	3	10	
Instance:		1	3	6	8	9	1	3	6	8	9	1	3	6	8	9
WEZ	A..	33	10	38	25	30	38	7	39	26	25	44	5	41	25	28
	B..	19	35	19	15	16	30	12	28	21	21	34	10	32	20	25
	C..	36	21	34	27	28	26	37	25	19	22	34	42	29	22	24
	D..	21	33	22	19	21	28	22	25	21	22	35	31	29	19	21

Table XVII lists the number of correct responses to positive instances of Pog, Mib, and Wez on each of tests A, B, C, and D. Turning first to Pog, results reveal that recognition was confined almost solely to instance 1 in Test A but dropped somewhat and spread to other instances upon succeeding presentations. This trend may be explained by the close similarity of the first teaching example and test instance 1. Both are unique in displaying a horizontal line

and a short arc, the only difference being in the relative position of the two elements. In all groups a reduction in response to instance 1 followed presentation of the second teaching example and was accompanied by a more accurate recognition of instance 7, particularly when assisted by the second positive teaching example. Here again similarity between teaching and test instances is in the form of unequal straight lines, thereby suggesting that a few individuals may have concentrated equally upon elements and relations, or even upon elements alone. The comparatively poor response to instance 2 and 4 (elements intersecting) until relieved by the third positive teaching example shows that "intersection" was a source of distraction to some.

In the case of Mib test instance 1 was more readily identified than were the other positive instances. This reaction was most pronounced following presentation of the first and second positive teaching examples, both of which display a circle touching on the <u>outside</u> of a square, in common with the test instance. The third positive example (circle <u>within</u> a square) lowered response to instance 1, but favored instances 4 and 9 in which one element is enclosed <u>within</u> the other. It appears, therefore, that the presence of this extra relationship of "insidedness" and "outsidedness" in some cases had a share in monopolizing attention and so delimiting perception of the relationship defining Mib. Undoubtedly test instance 6, the most radical

of the series in its departure from the teaching examples, was also the most difficult. But failure to identify this instance suggests that perhaps response was also to the figure or pattern as a whole, and that accentuation of the relation "larger than, smaller than" in such a case constituted a distortion.

In Wez are found some of the properties described for Mib. As before, those instances bearing closest resemblance to the immediate teaching example in point of relation, size, and shape elicited the greatest response. In Test A those positive instances which, like the teaching example, contain a circle on the <u>outside</u> of a triangle, were readily recognized, but not so with instance 3, in which the circle is <u>inside</u> the triangle. Difficulty with the latter was greatly alleviated by the introduction of the second positive example displaying a circle <u>inside</u> the triangle. The dominance of this relationship of "insidedness" and "outsidedness" is reemphasized by the fact that initial recognition of instance 3 was usually extended to include negative instance 5 (circle <u>inside</u>, but <u>not</u> touching the triangle).

Much of what has been saidcof Mib and Wez finds repetition in the results for Tov. With Zif, however, there is no question of "insidedness" or "outsidedness", and all positive instances evoke similar response. But one of the negative test instances, namely instance 10 (circle touching

both sides and an end of the rectangle), was of unparalleled difficulty. In all groups there was a better-than-chance tendency to regard this instance as a Zif. Inability of the second negative teaching example (circle touching one side and an end of the rectangle) to shift the response shows that the necessary relationship was only partially perceived. In the light of the teaching examples used, the validity of this particular test instance is questionable, for the former make it clear that a circle must touch both sides of a rectangle but they fail to specify that these must be the only two points of contact.* When it is remembered that instance 10 represents one quarter of the total number of negative test instances for Zif, the discrepancies in Table XVI are more easily understood.

Analysis also discloses that thexdifficulty experienced by many boys in Group I in identifying negative instances of Mef sprung from their loosely defining Mef as "a circle, partly black, partly white", and rejecting only those three instances (4, 7, 10) in which these qualities were absent. But with the advent of the first negative instance (circle, partly black, partly white) in Groups II and III, this hypothesis suffered a set-back.

An exhaustive analysis calls for a study of reaction to all test instances, both positive and negative.

---

* The fact that deduction of the rule governing Zif is impossible from positive examples alone constitutes an argument in favor of the use of negative examples.

But the foregoing, coupled with a further inspection of Table XVII yields the following tentative conclusions:

A. Negative instances are more easily identified than positive instances.

B. Item-difficulty is largely an inverse function of the similarity between teaching and test instances. This similarity effect is most apparent when the particular teaching and test instances are in juxtaposition, but diminishes somewhat upon interference by succeeding teaching examples.

C. The presence of relations incidental to the concept impedes solution. Items in which the definition is fulfilled irrespective of whether or not one of the elements involved is enclosed within the other are more difficult than items in which these added relations are an integral part of the necessary or defining relation. This special tendency may derive from the use of Dax as a demonstrative example.

D. Opportunity for hypothesis appears as a factor in item-difficulty. This conclusion supports Tyler's contention.[1]

E. The value of positive and negative teaching examples varies with different test instances within a given item. A particular example may assist one subject

---

1. Tyler, F. T. op. cit.

but not another on a given test instance, or it may help a subject with onextest instance but hinder him with another.

If little else, these results clearly depict how serious would be the consequences of a rearrangement or reconstruction of the teaching and test instances upon item-validity values. At the same time, they provide clues toward the improvement of validity in a number of cases.

4. Another Aspect of Validity.

It has been pointed out that the determination of whether a test is fulfilling the purpose for which it was constructed must consider not merely the validity of the test items, but the degree of correlation between those items. Thus, other conditions being satisfied, the lower the correlations, the higher the validity of the test as a whole[1]. To supplement the material of the preceding sections, Table XVIII lists intercorrelations of successively presented items for boys in Groups I and III. From the standpoint of validity it is apparent that the values for Group I most nearly fit the demands for a low intercorrelation of test items.

A careful study of the coefficients found in this latter group discloses several interesting facts:

---

1. Long, J.A., and Sandiford, P., op. cit., p. 119.

TABLE XVIII. INTERCORRELATION OF COMBINED A, B, C, AND D SCORES ON TEST ITEMS.

	GROUP I BOYS		GROUP III BOYS	
	r	SE_r	r	SE_r
MEF-VEC.....	.39	.13	.45	.12
VEC-MIB.....	.16	.14	.34	.13
MIB-ZUM.....	-.04	.15	.50	.11
ZUM-TOV.....	.26	.14	.30	.14
TOV-POG.....	-.08	.15	.28	.14
POG-WEZ.....	.11	.15	.52	.11
WEZ-ZIF.....	.55	.10	.67	.08
...				
VEC-ZUM.....	.55	.10		
ZUM-ZIF.....	.43	.12		
TOV-MIB.....	.14	.15		
MIB-POG.....	.09	.15		

A virtual or significant relationship existed between performances on easier items, but there was little or no connection between performances on the more difficult items. Also, scores on easy and difficult items were practically unrelated. One possible explanation for these low correlations is that scores on the difficult items may have been more dependent upon chance factors in contrast to scores on easy items. Again, the influence of transfer cannot be entirely ignored.

Ruger[1] in his experiment with mechanical puzzles was well aware of the significant effect of order of presentation upon problem-solving. He observed that some subjects were apt to generalize from one item to another as regards some detail of similarity which actually had no bearing upon its solution. Such behavior might explain the negative intercorrelation of Zum and Tov in Group I. Because of the large number of perfect scores in Zum, there was probably a distinct tendency to carry over the perceived relationship of "insidedness" and "outsidedness" to an attempted solution of Tov. Similarly, transfer effects, either positive or negative, may partly account for the presence or absence of interrelationships among the other test items.

5. Summary of Chapter V.

1. In the case of whole groups (N = 45) the test items were all of less than 50% difficulty, with errors ranging from 12 percent to 48 percent of the possible score.
2. Order of item difficulty correlated highly with diagnostic value, the most difficult items possessing a lower diagnostic value than the easier items.

---

1. Ruger, H.A., "The Psychology of Efficiency", Teachers College Educational Reprints, No. 5, 1926 (a reprint of Archives of Psychology, No. 15, 1910) pp. 29-30.

3. Positive test instances presented greater difficulty than negative instances. Other factors upon which item difficulty appeared dependent were
   (a) the extent of similarity between teaching and test instances.
   (b) the presence of extraneous relations within the stimulus pattern.
   (c) the number of likely methods of solution which suggested themselves.
4. The intercorrelations of test items in Group I Boys were inclined to be low and negligible; those in Group III Boys were considerably higher. In Group I easier items were significantly interrelated, while difficult items correlated low both with one another and with easier items.

CHAPTER VI.

GROUP REACTION TO SUCCESSIVE TEST PRESENTATIONS

Up to this point attention has been given over almost exclusively to a study of general test performance and the overall effectiveness of the several instructional methods. To estimate more fully the efficacy of the negative teaching example, the preceding study must be succeeded by a detailed analysis of step-by-step performance and an inquiry into the relative merits of the different modes of presentation as they affect perseveration and the formulation and rejection of hypotheses. Was progress toward solution gradual or rapid? How closely did progress on one item parallel that on another? With what degree of consistency did individuals respond to instruction by negative examples? This chapter will be devoted to answering these and other similar questions.

1. Improved and Unimproved Scores.

One approach to a study of the development of group reaction to repeated presentations of test stimuli is through a numerical consideration of improved and unimproved scores within each group. Unimproved scores, assembled in Table XIX, are subdivided three ways to include instances of fluctuating unimproved scores, reversals of judgment,

TABLE XIX. NUMBER OF UNIMPROVED SCORES OUT OF A POSSIBLE 120 FOR EACH SUB-GROUP, CLASSIFIED IN TERMS OF DECLINATION[1], REVERSAL OF JUDGMENT[2], AND PERSEVERATION[3]

	GROUP I								GROUP II								GROUP III							
	Boys				Girls				Boys				Girls				Boys				Girls			
	H	M	L	Total	H	M	L	Total	H	M	L	Total	H	M	L	Total	H	M	L	Total	H	M	L	Total
DECLINATION	27	43	44	114	28	40	42	110	28	39	43	110	31	33	44	108	30	38	32	100	21	28	32	81
REV.OF JUDG.	2	3	1	6	3	2	1	6	3	3	2	8	6	5	8	19	3	3	6	12	1	5	4	10
PERSEVERATION	18	8	5	31	15	5	14	34	9	6	7	22	8	8	5	21	8	7	2	17	19	13	6	38
TOTAL				151				150				140				148				129				129
TOTAL PERCENTAGE OF UNIMPROVED SCORES				42				42				39				41				36				36

1. includes all fluctuating scores which fail to improve beyond the A score.
2. a perfect A score terminating in an imperfect D score.
3. an unchanging score which is not a perfect score.

and perseveration, each defined as in the table.

Of a total of 360 items, 36 to 42 percent displayed a lack of improvement in attempts beyond the initial or A Test, though in 96 percent of all such cases A-scores were of 50% grade or better. Results were inconclusive in ascribing greater progress to one method than to another, though there is some evidence that the absence of memory hastened improvement among low IQ groups.

According to further observation, reversale of judgment were relatively infrequent. Perseveration occurred in a greater number of cases and offered limited support for the theory that the negative teaching example tends to interfere with mental inertia. However, because of the restricted definition of perseveration, imposed by the very nature of the experiment, any further conclusions would be misleading. For example, indications that the high groups were equally or more often subject to mental inertia than other groups quite overlooks the fact that perseverative response by the former normally occurred on a higher score level, the only obstacle to a perfect score sometimes being a probable perceptual oversight or a flaw in the test itself. This fact leads to an inescapable admission of the manifold difficulties besetting an objective analysis of this type of behavior.

Isolation of perseveration in all its aspects entails a recognition of such possible forms of reaction as unswerving response to detail or to the whole pattern, concentration fixed upon similarities or upon differences, unyielding emphasis upon elements rather an upon relations. Yet it is very doubtful if more than a minority of such cases could be covered by an account which regards unaltered responses to individual test instances as the sole outward manifestations of perseveration.

2. Perfect Scores.

Group progress may be further analyzed by a consideration of perfect scores. Reference has already been made to instances embodying reversals of judgment wherein a perfect A scofe is coupled with a subsequent decline in achievement. Table XX furnishes additional data and provides for more extensive conclusions, These may be stated in brief:

1. Of all scores displaying improvement beyond the initial or A score, 18 to 28 percent were solutions in the sense in which this term applies to perfect D scores.
2. Under successive presentation, the negative teaching example appeared to have a neutral, if not reductive effect upon the number of solutions achieved.

TABLE XX. NUMBER OF ITEMS SOLVED* BY SUB-GROUPS TOGETHER WITH THE TOTAL NUMBER OF SOLUTIONS OUT OF A POSSIBLE 360 FOR WHOLE GROUPS.

	GROUP I		GROUP II		GROUP III	
	Boys	Girls	Boys	Girls	Boys	Girls
HIGH	24	24	21	22	29	30
MEDIUM	16	14	12	11	20	18
LOW	17	6	8	6	8	17
TOTAL	57	44	41	39	57	65

3. Both boys and girls solved more concepts under cumulative than under successive presentation involving the negativexexample.

4. High groups were credited with more than 40 percent of all solutions and gave evidence of a better-than-chance tendency that a perfect score occurring in Test A would maintain itself throughout all tests.**

3. Mean Scores.

Probably the most adequate method for estimating the extent of group advancement beyond the initial test is by a comparison of average A, B, C, and D scores. The "learning" curve described for each of the major groups (Table XXI) indicates that A scores were little battered in succeeding tests, the gain nowhere exceeding ten

---

* perfect D score.
** Appendix II, Table E.

TABLE XXI. GROUP PERFORMANCE ON TOTAL OF EACH OF TESTS A, B, C, AND D. MEAN SCORES, STANDARD DEVIATIONS, AND STANDARD ERRORS.

		GROUP I		GROUP II		GROUP III	
		Boys	Girls	Boys	Girls	Boys	Girls
TEST A	A.M.	50.90 (.93)	50.19 (1.06)	49.21 (1.13)	51.34 (1.05)	51.17 (1.40)	53.48 (1.01)
	S.D.	6.24 (.66)	7.08 (.75)	7.60 (.80)	7.04 (.74)	9.40 (.99)	6.80 (.72)
TEST B	A.M.	53.12 (.88)	52.68 (1.15)	45.83 (1.77)	43.88 (1.79)	48.41 (1.83)	54.46 (1.24)
	S.D.	5.88 (.62)	7.68 (.81)	11.84 (1.25)	12.0 (1.26)	12.24 (1.29)	8.28 (.87)
TEST C	A.M.	53.30 (1.06)	53.12 (1.15)	53.74 (1.25)	55.43 (1.31)	53.83 (1.63)	58.90 (1.58)
	S.D.	7.12 (.75)	7.68 (.81)	8.40 (.89)	8.80 (.93)	10.90 (1.15)	10.56 (1.11)
TEST D	A.M.	53.57 (1.06)	52.86 (1.25)	48.50 (2.01)	47.52 (1.98)	51.43 (2.15)	57.12 (1.41)
	S.D.	7.08 (.75)	8.40 (.89)	13.48 (1.42)	13.24 (1.40)	14.40 (1.52)	9.48 (1.00)

percent of the A score. In general outline, developmental reaction assumed some of the characteristics displayed by Dickinson's groups[1]. For example, Group I demonstrated most gain from Test A to B, with improvement thereafter becoming almost imperceptible. And in Groups II and III performance was marked by decided fluctuations. Immediately

1.

following presentation of the first negative example there occurred a drop in response, attended by an increase in variability. In Test C scores rose sharply and displayed greater uniformity, only to suffer a relapse in Test D. Of these groups only girls in Group III failedcto conform to this general pattern.

Group progress has been measured statistically by considering vertical differences between the mean scores of Table XXI. These differences, interpreted in terms of critical ratios (Table XXII) indicate a most "rapid" change in scores from Tests B to C for Groups II and III, the drop in scores thereafter displaying significance only within Group II. Another noteworthy fact is that this latter group in its progress from A to C actually surpassed by a small margin the progress achieved by Group I.

Measurement of progress of dull and bright subgroups* calls for an amendment to earlier conclusions which stressed the unfavorable effect of the negative teaching example upon low-group achievement. Thus, a comparison of differences credits the system of positive-negative presentation with promoting the greatest gain from A to C among low groups. Or to state it in a different way: There is some evidence that for these groups positive teaching examples are generally more effective when

---

* Appendix II, Tables G to J.

TABLE XXII. CRITICAL RATIOS OF THE DIFFERENCES* BETWEEN MEAN TEST SCORES WITHIN EACH GROUP (BRACKETED LETTERS DESIGNATE TEST HAVING HIGHEST SCORE)

	GROUP I		GROUP II		GROUP III	
	Boys	Girls	Boys	Girls	Boys	Girls
A and B	2.20(B)	2.62(B)	2.24(A)	4.78(A)	1.66(A)	1.54(B)
B and C	.23(C)	.70(C)	4.88(C)	6.80(C)	3.76(C)	4.04(C)
C and D	.48(D)	.39(C)	3.03(C)	4.32(C)	1.63(C)	1.58(C)
A and C	2.82(C)	3.02(C)	4.19(C)	3.47(C)	2.31(C)	4.34(C)
A and D	3.18(D)	2.49(D)	.39(A)	2.11(A)	.13(D)	2.91(D)
B and D	.59(D)	.29(D)	2.59(D)	4.04(D)	2.96(D)	3.59(D)

* For differences in variability see Appendix II, Table F.

immediately anteceded by a negative example than when given in continuous series with positive examples. With the high groups improvement from A to C was more pronounced under the influence of positive successive presentation as against positive-negative successive presentation, suggesting that in this case positive examples may possibly have greatest value when immediately preceded by other positive examples rather than by negative examples. A precise measurement of these effects was difficult to obtain for the reason that the positive examples studied immediately prior to Test C in groups I and II were not identical, the second teaching example in the former group serving as the third example in the latter group. Finally, differences between A and D Test scores, though not statistically computed, were of

GROUP PROGRESS ON INDIVIDUAL TEST ITEMS

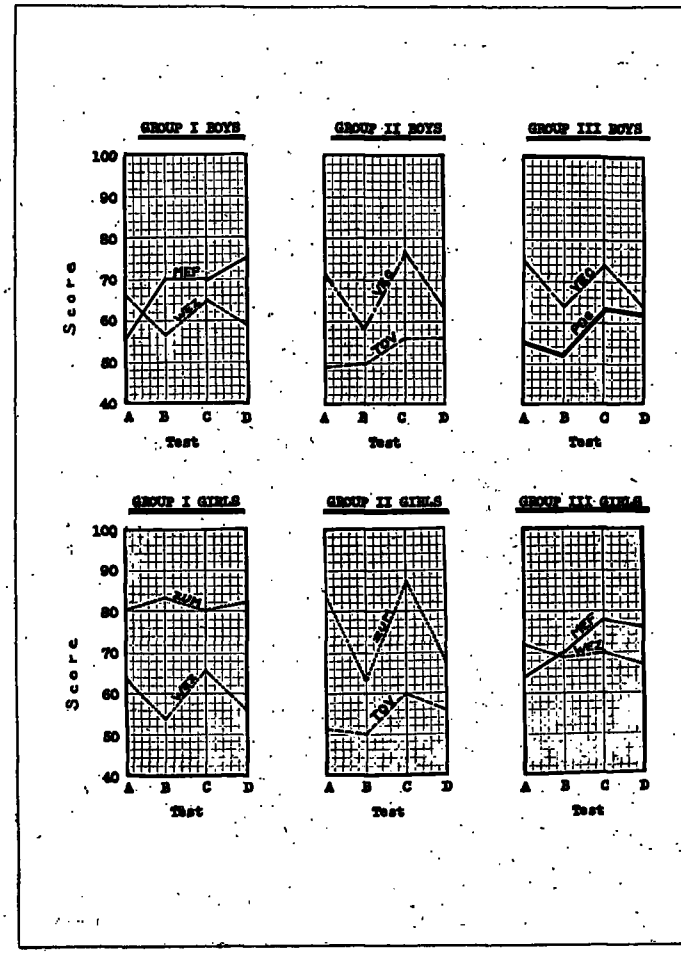

Fig. 11

sufficient size to suggest that low groups progressed most "rapidly" when faced with cumulative presentation; of the high groups the girls gave evidence of greatest progress from Tests A to D under cumulative presentation, while the boys achieved the greatest advance under positive successive presentation.

In order to comprehend more fully the nature of the "learning" gradient, graphical outlines of progress on individual test items were drawn for each Group (Fig. 11) revealing a number of opposing trends in performance. For instance, concerning boys in Group I it was found that the second positive teaching example was immediately followed by a sharp rise in Mef scores and an equally sharp decline in Wez scores, with a repetition of such performance succeeding presentation of the fourth positive example. The contrasting effects of different teaching examples were also reflected in Group I girls' scores for Zum and Wez. With the boys in Group II a sharp drop in Vec scores immediately accompanied presentation of the negative examples, in opposition to the imperceptible changes in corresponding Tov scores. The "learning" curves for Group II girls likewise indicate that some of the easier items were characterized by greater fluctuations in performance than were the most difficult items. Turning to a consideration of Group III boys, a decline in achievement from Test A to D was associated with one of the easiest

items, namely Vec, whereas considerable progress was made on one of the most difficult items, Pog. In the girls' group a similar difference in trend occurred in two items of almost equal difficulty.

This lack of parallelism between performance on test items prompts an inquiry into how closely total performance on one test was associated with that on another. Table XXIII reveals a significant relationship between test scores, with correlations varying from near low to high. Among

TABLE XXIII. INTERCORRELATIONS OF AVERAGE A, B, C, AND D TEST SCORES.

	GROUP I		GROUP II		GROUP III	
	Boys	Girls	Boys	Girls	Boys	Girls
A and B	.38	.63	.53	.50	.50	.53
B and C	.68	.85	.47	.43	.66	.72
C and D	.86	.85	.52	.44	.73	.72
A and C	.64	.62	.59	.52	.72	.61
A and D	.65	.58	.42	.41	.44	.51
B and D	.71	.87	.86	.89	.88	.85

conclusions that may be drawn are the following:

1. Initial performance afforded but a rough estimate of final achievement.
2. Average individual reaction to the first and second negative teaching examples displayed a high degree of consistency.

3. Performance aided by positive-negative successive presentation appeared slightly more irregular and unpredictable than when assisted by either of the two remaining methods, as suggested by the lower correlations in Group II.

4. Positive and Negative Component Scores

In the last chapter brief reference was made to the accuracy with which negative test instances of a concept were identified. A broader picture of group performance in this respect is afforded by Table XXIV, which converts average positive and negative component scores into percentages for high, medium, and low I.Q. groups. It expresses an unmistakeable tendency by these groups to score higher on every test in their recognition of negative instances, the only two exceptions attaching to C and D scores of High Group I. Also observable is the disparity between high and low group accuracy in responding to positive instances and the closer similarity of their negative scores. Comparisons between high and low groups based on the above table show that in 19 out of 24 cases the difference between positive component scores was fifteen points or more, to the disadvantage of the low group; but in 17 out of 24 cases low group negative scores

TABLE XXIV. PERCENTAGE ACCURACY OF POSITIVE AND NEGATIVE COMPONENT SCORES ACHIEVED BY WHOLE AND SUB-GROUPS ON TESTS A, B, C, AND D.

		GROUP I				GROUP II				GROUP III			
		Boys		Girls		Boys		Girls		Boys		Girls	
		+	−	+	−	+	−	+	−	+	−	+	−
TEST A	H	57	70	57	76	57	74	57	78	61	80	61	80
	M	48	72	41	80	43	74	46	81	48	81	46	81
	L	48	80	35	81	35	78	43	74	41	65	52	76
	ΣHML	50	74	43	80	46	76	48	78	50	76	52	80
TEST B	H	65	72	61	76	57	72	54	78	57	78	65	83
	M	50	76	46	85	35	70	33	65	43	72	43	80
	L	48	81	37	81	39	65	26	65	37	67	50	80
	ΣHML	54	76	48	81	41	69	37	69	46	72	52	81
TEST C	H	74	69	65	78	67	76	63	80	70	76	76	80
	M	52	74	46	83	54	81	57	81	57	81	57	80
	L	52	76	41	80	37	81	54	76	52	67	65	81
	ΣHML	59	74	50	80	52	80	57	80	59	74	65	80
TEST D	H	72	69	63	80	67	74	59	80	63	78	74	83
	M	50	80	43	85	37	70	39	69	52	76	50	80
	L	50	80	35	83	39	69	37	69	46	67	59	80
	ΣHML	57	76	48	83	48	70	43	72	54	74	61	81

were greater than or no more than five points below high group negative scores. In other words, the low IQ groups fell far behind in the detection of positive instances, but demonstrated closer ability with the high groups in handling negative instances, especially in Group I where they actually

surpassed the high groups in this respect. The medium groups, instead of retaining a position midway between these two extreme groups, behaved more nearly like the low group. This line of demarcation distinguishing the performance of high groups from that of medium and low groups suggests that it is not altogether impossible that there may be a point along the intelligence scale at which the ability to score reasonably high in the identification of positive instances suddenly makes itself felt. There is need for further experimentation of this nature, involving larger and more representative groups.

Several studies have examined the effect of positive and negative instances insofar as they concern the simple recognition of different materials. Achilles[1], experimenting with geometric forms, words, nonsense syllables, and such like, first presented to his subjects a number of items and then required that they respond to a recognition test comprised of these and numerous new or unfamiliar items. It wasffound that greater accuracy characterized response to the unfamiliar than to the unfamiliar, but there is no record of the statistical reliability of this trend. The author concluded that "the new make a distinct impression and the subject responds with more certainty.....This strangeness or

---

1. Achilles, E.M., "Experimental Studies in Recall and Recognition", Archives of Psychology, vol. VI, Sept. 1920, pp.

newness appears to be a positive thing." A similar experiment by Seward[2] involved the presentation of a series of papers bearing different designs and colors, followed by an interpolated task, and then the recognition test. The correctness of the positive response was observed to be directly proportional, that of the negative response inversely proportional to the degree of identity between the presentation and the test stimuli. That is, identity between original and immediate stimuli gave rise to a more accurate positive response, while dissimilarity between the two favored the negative response. Tendency differences were regarded as being highly reliable for the select group used; correlations with intelligence were inconclusive.

The close congruency between these results and those issuing from the present study lists the possibility that the confronting problem, oversimplified by some, may have resolved itself into one of mere recognition. Born of a misunderstanding of the preliminary instructions, there may have developed a strong tendency to seek in a test facsimiles of the teaching examples, with an eye to exactness of size, shape, position, and number of elements. Because of the wide diversity of detail between most teaching and test stimuli, with the resultant emphasis upon dissimilarity rather than upon similarity, subjects

2. Seward, G.H., "Recognition Time As A Measure of Confidence (An Experimental Study of Redintegration)", Archives of Psychology, Vol. XVI, 1928, pp. 1-54.

may have been more readily able to identify negative than positive instances. This theory finds a measure of support in material of the preceding chapter.

5. Summary of Chapter VI.

1. A majority of scores showed some improvement beyond Test A.
2. Testing conditions made difficult a complete and objective analysis of perseveration.
3. Accurate prediction of final achievement was impossible on the basis of initial performance.
4. The apparent inhibitory effect of the negative teaching example, manifested by an adverse change in both mean score and variability, was immediate rather than of prolonged duration.
5. Negative teaching examples were highly consistant in their over-all immediate effect upon individual performance.
6. Among low groups there were indications of the greater effectiveness of positive examples when immediately preceded by negative examples. For the high groups the positive teaching example seemed most effective when preceded only by other positive examples at least as far as successive presentation was concerned. (These tendencies were not statistically reliable).

7. The negative teaching example failed to augment the number of complete generalizations and may actually have hampered their development.
8. The study of positive and negative component scores offers a practical approach to performance-analysis.

CHAPTER VII.

FINAL EVALUATION AND CRITICISM

The contrasting results of the various experimental studies in concept formation underscore the supreme importance of reducing and removing the possibility of inaccuracies inflicted by the presence of uncontrolled variables of one kind or another. For the successful control of these variables in individual experiments more adequate facilities are at hand than in group experiments where increased complexity adds to the difficulty of the situation.

Oneo of the chief prerequisites in the experimental control of test performance is a complete and comprehensive set of instructions. Without this proper guidance the isolation and measurement of special abilities becomes virtually impossible, for, as Thurstone[1] writes, "the fact that a person has a high rating in a particular ability does not help him to superior performance in a task unless the task involves the ability in question." If, in these tests, language deficiencies and a consequent inability to take full advantage of the instructions obstruct the most complete expression of generalizing ability of which the

---

1. Thurstone, L.L., op. cit., p. 3.

subject is capable, then test validity is seriously impaired. In Wood's experiment this difficulty was countered by a certain flexibility[1] in the instructions which permitted their adaptation to individual needs. Furthermore, the essential difference between positive and negative examples was thoroughly "stamped in" by having the subjects place each in separate piles, thus supplementing visual and auditory with kinaesthetic cues. These arrangements provided at the outset a reasonable assurance of the subject's full acquaintance with the demands of the task before him to a degree not possible in group experiments such as the present one. Therein probably lies one of the principal reasons why the present results differed so markedly from those obtained under a system of individual testing.

In other words, there is the possibility that generalizing ability, as related to non-verbal material, was one among several abilities evoked by the present group tests, and that, if so, inadequacy of the instructions may have been partly responsible for this state of affairs. This supposition finds some basis in the high percentage of unimproved scores. Also, the sharp drop in average scores on Tests B and D, coupled with the high degree of relationship between such performance, is supporting

---

1. Delivery of the instructions concluded with the words, "If you do not understand any part of what you are to do, please ask me about it now." Wood, J., op. cit., Appendix.

testimony that the value of the negative example was appreciated by only a limited number. This may signify lack of proper guidance or it could mean that the negative example was detrimental in itself. The probability of the former is suggested by a careful re-examination of the directions accompanying the tests in the light of various findings, a matter which later will be discussed in greater detail.

Of course, differences between experimental results cannot be justified in terms of differences in test administration alone. In analyzing reaction to individual test instances in an earlier chapter, it was observed that the whole character of the tests could be made to undergo considerable change by slightly altering or reshuffling the test material. Because practical necessity in the present case demanded that such changes be made in Wood's test material, already a modification of Smoke's original, our tests may have greater or lesser potentialities for measuring generalizing ability at a given age level than those tests from which they were constructed. All comparisons must take into account this fact; especially is this true where Dickinson's experiment is involved, for here it was deemed advisable to carry these changes even further by completely redefining several test items to fit the needs of a still younger group.

One need hardly continue further to realize that it is difficult enough to classify one form of a test as "more valid" or "less valid" than some other, but much more difficult to makexa pronouncement as to its ultimate validity. The problem must be subject to attack, not from one fixed point of view, but from all sides. The danger of resting judgment upon mere statistical formulae in analyses of this sort is cited by Kuhlmann who charges that statistical method "puts its main faith in the possibility of what in effect amounts to correcting error made in observations after they have been made, of supplementing or supplying observations where none exist."[1] For example, application of the index of reliability to determine the validity of the tests under consideration would bestow upon them high values which are unsubstantiated by the results of a more extended examination. This possibility went largely unnoticed by Dickinson[2] who concluded that her tests were valid solely on the basis of the index of reliability and a highly subjective analysis of the process of concept formation.

And so, in reviewing present findings associated with test validity, the only deductions that can be reliably made must take the form of recommendations for

---

1. Kuhlmann, F., "Our Changing Fashions in Methods of Research", American Journal of Psychology, vol.55, 1942, p.572-3.
2. Dickinson, A.E., op. cit., p. 51-2.

the more effective control of test procedure. Concerning the actual validity of these group tests, a final verdict must await further experiment, for a close study of individual performance and of positive and negative component scores has made it appear not unlikely that recognitive ability rather than generalizing ability was frequently being tested.

Duda p. 69.

CHAPTER VIII.

CONCLUSIONS, IMPLICATIONS, AND SUGGESTIONS FOR FUTURE RESEARCH

1. Conclusions

The expanding role of conceptual thinking in human endeavor pointed to the need of developing tests for the accurate measurement of such ability. Concept formation was used synonymously with generalization, and was defined according to Smoke's usage of the term as "a process whereby an organism develops a symbolic response (usually but not necessarily linguistic) which is made to the members of a class of stimuli patterns, but not to other stimuli."[1] It is primarily a process of responding to common relationships, though elements would appear to constitute a necessary part thereof. A brief outline of relevant studies suggested that more repetition and continuation of previous experiments would help satisfy a need for the perfecting of techniques and the establishment of a basis for more extensive generalizations. Accordingly, it was decided to check hypotheses advanced by previous experimenters utilizing Smoke's technique of guaging conceptual ability in terms of the ability to perceive an inter-element relationship common to a series of geometric patterns. Since Wood[2] had already applied this technique to the individual study of generaliz-

1. Smoke, K. L., op. cit.; p. 8.
2. Wood, J. A., op. cit.

ing ability in Grade VI boys, its application to a group study at the same level of educational attainment appeared worthy of investigation.

To permit analysis of the varied effects of instruction upon success in generalizing, arrangements were made to study performance under three sets of conditions. For this purpose, subjects were selected and matched with one another according to sex, chronological age, and I.Q. to form three experimental groups (exclusive of trial groups), each comprising 45 boys and 45 girls. These in turn were subdivided into groups representing children of high, medium, and low intelligence. The general procedure required the presentation, by means of film slides, of a series of teaching and test instances for nine different concepts. The study-time for each of the four teaching examples was 8 seconds, while the total time required for response to the 10 test instances was 25 seconds. The three experiments were alike in their use of a fore-test and all employed the same tests, each made up of an almost equal number of positive and negative instances of a given concept; they differed, however, in regard to the type of teaching examples employed and to their manner of presentation. The first group was subjected to instruction by the successive presentation of positive examples; the second and third groups were instructed by means of both positive and negative examples, involving successive presentation and cumulative presentation, respectively.

Cumulative presentation, as opposed to successive presentation, provided for the continued exposure of the examples during the period of testing. The test was taken immediately following the study of each teaching example, the four presentations of the same test being designated as A, B, C, and D. A set of standardized directions accompanied test administration in all groups.

Owing to the high correlations which defined the interrelation of the total of D scores with the sum total of A, B, C, and D scores, the former was regarded as a suitable criterion of test performance. Among the more important findings of this study were the following:

1. These tests are capable of group measurement with a reasonably high degree of reliability.
2. Generally speaking, test performance under the influence of successive presentation was more satisfactory where only positive examples were employed. Where both positive and negative examples were involved, cumulative presentation appeared the better method.
3. While girls were credited with maximum achievement, there was no conclusive evidence for the existence of sex differences.
4. Test performance, while showing some positive relationship to intelligence, reading, and arithmetic reasoning, appeared also to be measuring an ability

or abilities beyond the scope of these classifications. Test performance seemed most closely associated with scholastic achievement.

5. Negative teaching examples, to presentations of which average individual reaction displayed high consistency, were usually accompanied by an immediate decline and spread in group achievement.
6. A comparison of high and low I.Q. groups in the basis of all four test performances suggested that the negative example was of little advantage to bright children, while a handicap to those of more or less average intelligence. Among the latter, on the other hand, there were indications that the negative teaching example enlivened and intensified the didactic effect of the positive example immediately following. This effect was not duplicated in the case of the high groups.
7. The value of the negative teaching example varied with the individual and with the particular test instances employed.
8. Negative test instances were identified by all groups with greater accuracy than were positive instances. Low I.Q. groups demonstrated close ability with high I.Q. groups in identifying negative test instances, but were much less capable in regard to positive instances.

9. Item-difficulty was contingent upon -

   a. The similarity and dissimilarity between teaching and test instances.

   b. The presence of relations within the stimulus pattern which were irrelevant to a solution.

   c. The number of approaches which appeared likely to lead to a solution.

10. Analysis of test validity was confined to two boys' groups, the one instructed by successive presentation of positive examples, the other by cumulative presentation of positive and negative examples, and revealed that -

   a. Order of difficulty was closely related to the diagnostic value of an item, the easier items differentiating more effectively between able and poor performers.

   b. Intercorrelations of test items were generally low and negligible where performance was uninfluenced by the negative example, but were higher for the group instructed by positive-negative cumulative presentation.

   The results caution against too great reliance upon any one statistical formula or technique for the determination of test validity.

2. Educational Implications

The implications of this study for Educational or Applied Psychology may be briefly summarized: Where conceptual thinking is involved at the Grade VI level, use of the negative example in group instruction is apt to be more confusing than beneficial unless painstaking care is exercised.

In many cases it is probably not so much the negative example itself which provokes confusion, but rather an erroneous conception of the problem to be solved. For instance, to grasp the full import of the negative example, one must first understand the significance of the positive example; and to understand the positive example requires, in the present tests, awareness that some sort of relationship is involved. It would appear, therefore, that the value of the negative example is governed not only by a familiarity with the demands of the problem but also by the type of material which forms the object of generalization.

3. Suggestions for Future Research

As often pointed out, the value of many a psychological investigation has been sacrificed by the all-too-frequent tendency to abandon a project at a certain stage of development and before some practical and worthwhile contribution to knowledge has been realized. In the study of generalizing ability (concept formation) recent attempts, notably those of Long and Welch, have sought to remedy this situation. In keeping with this trend, a set of similar studies is being currently conducted at the University of British Columbia, of which the present one is the third in the series. As an inducement toward the continuation of this endeavor, the following suggestions and recommendations are offered for the improvement of testing techniques:

Foremost among the factors demanding revision is the preliminary guidance which is intended to introduce the subject to the problem situation. In remodeling the instructions special emphasis should dwell upon three objectives:

1. The Subject must be impressed with the fact that relations are involved, and relations only.
2. He must understand the difference between positive and negative teaching examples, and the significance of each.
3. He must understand that all positive teaching examples of a given item contain one relation in common and must be warned that no one example is unique in this respect.

As a first step toward accomplishing these ends, Dax might well be replaced by another "concept" which, after the pattern of Vec, does not involve a closed figure. The effect of this substitution upon the solution of items which include "insidedness" or "outsidedness" as either incidental or essential relationships could then be analyzed and comparisons made. In this way the influence of the fore-test upon subsequent performance and item-difficulty can more easily be judged.

If these tests are to provide even a rough measure of generalizing ability, great care must be applied in formulating the instructions. The importance of this requirement can not be exaggerated. Among several changes that should be made in the present set of directions, at least two of these bear mentioning. For example, accompanying the showing of the second teaching example, the words "we might

guess...that a Dax is a dot and a triangle" were intended to lead the subjects in their individual efforts toward the correct generalization, "a dot inside a triangle". But the possibility that this suggestion may also have misled thinking toward elements and away from relationships is not denied by the facts. Another shortcoming reflected in test performance is to be found in the part-statement that "....the position of the dot does not matter," referring to its position within the triangle. Even though the correct relationship is later defined, this statement is not sufficiently explicit and leaves too much room for confusion in the mind of the subject.

Keen judgment should govern the selection and arrangement of the teaching and test instances, in the interest of validity. All approach toward identity of like teaching and test instances in respect to shape, size, and position of elements should be avoided, particularly in the case of items of lesser difficulty.

In preparing the tests for administration at higher age levels, validity might be best served by an adjustment of the time factor rather than by changes in the test material itself.[1] Group response to positive and negative test instances should be studied, and any trends noted. In the event that response follows a pattern similar to that

---

1. Thurstone, T. G., op. cit., p. 335

observed in this study, the possibilities of the positive component score as a suitable criterion of generalizing ability should be considered by computing its relationship to other variables and by analyzing its diagnostic capacity.

Where possible it would be of interest to compute correlations between generalizing ability and intelligence as measured by an individual test such as the Stanford-Binet in which reading has a more limited role.

Dickinson has proposed the time-saving measure of deferring testing until all four teaching examples are presented, thus eliminating Tests A, B, and C. While the practical worth of such an arrangement is attested by the high inter-relationship of D scores with the total of A, B, C, and D scores, its adoption at this stage of development would hamper the study of validity and circumscribe all efforts to probe the true nature of individual and group performance.

The advantages claimed by Thurstone[1] for the projector method of test administration are, firstly, maximum control over exposure-time, and secondly, facility for capturing and holding attention. He points out that "the attention value of the visual projector method can be regarded as one of its principal features". To ensure that this statement applies in any given situation, care should be taken to minimize

---

Thurstone, L.L., "A Micro-Film Projector Method for Psychological Tests", Psychometrika, vol. VI, #4, August 1941, p. 240.

distractive influences by placing the projectors as far to the rear of the room as possible. Use of a portable screen frequently makes this impossible owing to its limited size. A better substitute would be a large white sheet or, if a portable screen must be used, the same effect could be produced by contracting the size of the slide-images or by employing a different type of projector-lens.

BIBLIOGRAPHY

Achilles, E.M., "Experimental Studies in Recall and Recognition", Archives of Psychology, vol. VI, Sept. 1920.

Billings, M.L., "Problem-Solving in Different Fields of Endeavor", American Journal of Psychology, vol. XLVI, 1934. pp. 259-272.

Dickinson, A.E., An Investigation Into The Generalizing Ability of Grade Two Pupils, Master's Thesis, Vancouver, University of British Columbia, 1943, published in abstract in Journal of Educational Psychology, vol. XXXV, 1944. pp. 432-441.

Ewart, P.H. & Lambert, J.F., "The Effect of Verbal Instructions Upon the Formation of a Concept", Journal of General Psychology, vol. VI, 1932. pp. 400-413.

Garrett, H.E., Statistics in Psychology and Education, Toronto, Longmans, Green and Co., 1940.

Hanfmann, E. & Kasanin, J., "A Method for the Study of Concept Formation", Journal of Psychology, vol. III, 1937. pp. 521-540.

Hull, C.L., "Quantitative Aspects of the Evolution of Concepts; An Experimental Study", Psychological Monographs, vol. XXVIII, No.1, 1920.

Jackson, R.B. & Ferguson, G.A., Studies on the Reliability Tests, Bulletin No.12 of the Department of Educational Research, University of Toronto,1941.

Kuhlmann, F., "Our Changing Fashions in Methods of Research", American Journal of Psychology, vol. 55, 1942. pp. 569-573.

BIBLIOGRAPHY (Continued)

Kuo, Z.Y., "A Behavioristic Experiment on Inductive Inference", Journal of Experimental Psychology, vol. VI, 1923. pp. 247-293.

Long, J.A. & Sandiford, P., "The Validation of Test Items", Bulletin No.3 of the Department of Educational Research, University of Toronto, 1935.

Long, L. & Welch, L., "A Preliminary Investigation of Some Aspects of the Hierarchical Development of Concepts", Journal Of General Psychology, vol.XXII, 1940. pp. 359-378.

McGeoch, J.A., Psychology of Human Learning, New York, Longmans, Green and Co., 1942.

Maier, N.R.F., "Reasoning in Rats and Human Beings", Psychological Review, vol. XLIV. 1937. pp.365-378.

Peterson, G.M., "An Empirical Study of the Ability to Generalize", Journal of General Psychology, vol. VI, 1932. pp. 90-114.

Ruger, H.A., "The Psychology of Efficiency", Teachers College Educational Reprints, No.5, 1926 (a reprint of Archives of Psychology, No.15, 1910).

Seward, G.H., "Recognition Time as a Measure of Confidence (An Experimental Study of Redintegration)", Archives of Psychology, vol. XVI, 1928. pp.1-54.

Sherman, M., Intelligence And Its Deviations, New York, The Ronald Press Co., 1945.

Smoke, K.L., "An Objective Study of Concept Formation", Psychological Monographs, vol. XLII, No.4, 1932.

_____ "Negative Instances in Concept Learning", Journal of Experimental Psychology, vol. XVI, 1933. pp. 583-8.

Symonds, P.M., "Factors Influencing Test Reliability", Journal of Educational Psychology, vol. XIX, 1928. pp. 73-87.

BIBLIOGRAPHY (Continued)

Thompson, J., "The Ability of Children of Different Grade Levels To Generalize on Sorting Tests", Journal of Psychology, vol. XI, 1941. pp. 119-126.

Thurstone, L.L., "Primary Mental Abilities", Psychometric Monographs, No.1, 1938.

_____ "A Micro-Film Projector Method For Psychological Tests", Psychometrika, vol. VI, No.4, August 1941.

Thurstone, T.G., "The Diccifulty of a Test and Its Diagnostic Value", Journal of Educational Psychology, vol. XXIII, 1932. pp. 335-343.

Tyler, F.T., Generalizing Ability of Junior High School Pupils: An Experimental Study of Rule Induction, unpublished Ph. D. Thesis, University of California, 1939.

Wolfe, D., "Factor Analysis to 1940", Psychometric Monographs, No.3, 1940.

Wood, J.E., The Relative Role of Positive and Negative Instances in Concept Formation, unpublished Master's Thesis, Vancouver, University of British Columbia, 1943.

Woodworth, R.S., Experimental Psychology, London, Methuen and Co. Ltd., 1938.

APPENDIX I.

A. Instructions Issued Subjects In Experiment I.

Today your teacher has suggested that you help us work out some picture-puzzles. We think you will enjoy doing these puzzles. You have never seen them before. To make it more interesting we shall score your results. Here on the black-board you see the first part of your answer sheet: Fill in your name, whether a boy or girl, your age, birthday, school, and the name of your teacher. Pay no attention to the other blanks.

First on the screen we are going to show you a picture of a thing called a Dax (spell). You will study this picture for a few moments to discover the idea of what a Dax is. Then you will be shown the puzzle made up of 10 pictures. You are to tell which of these 10 pictures contain the idea of what a Dax is, and which do not.

To make this clear, let us look at the first example. (Dax flashed on screen). Here is a picture of a Dax. Now study it and see if you can decide what a Dax is. (Pause. Then Dax replaced by test). Now look at the puzzle. Look at number one. Do you think it contains the idea of a Dax? If you do then draw a circle around "yes" in row A under "1". If you think it does not contain the idea of a Dax, circle the "no" in row A under "1".

(Illustrating) Then look at picture Number 2. Do you think it contains the idea of a Dax? If you do, then draw a circle around "yes" in row A under "2". If you think it does not contain the idea of a Dax, then circle the "no" in row A under "2". (Illustrating) Now you do the same for the other pictures. Draw the circles around "yes" or "no" in row A, because this is the first puzzle. Pay no attention to these columns on the right. Be sure that you put your answers under "Dax" on your sheet. (Pause) Raise your hand when you have finished. (Test replaced by second Dax).

Now let us look at the second example on the screen. This is also a Dax. Now you must remember what the first Dax was like, and see how this one is like the first one. You remember that the other example had a triangle and a dot. So has this one. Remember that the shape of the other triangle was not the same as this one, so that the shape of the triangle does not matter. You remember also that the dot in the other Dax was in a different position, so that the position of the dot does not matter. We might guess, then, that a Dax is a dot and a triangle. Now let us look at the second puzzle. (Dax replaced by test) Look at picture number one. Do you think it contains the idea of a Dax? If you do, then circle the "yes" in row B under 1, because this is the second puzzle. If you think picture number one does not

contain the idea of a Dax, then circle the "no". (Illustrating) Now look at picture number two. Do you think it is a Dax? if you do, then circle the "yes" in row B under 2. If you think it is not a Dax, circle the "no". (Illustrating) Now you do the rest of the puzzle. (Pause. Test replaced by third Dax)

Now let us look at the third examplexof a Dax. Do you think that a Dax is a triangle and a dot? Well, I am going to tell you what a Dax <u>really</u> is. A Dax is <u>a triangle with a dot inside it</u>. You remember that in each case the dot was <u>inside</u> the triangle. Now you do the third puzzle. (Pause. Then Dax replaced by test) Do you think picture number one is a Dax? Yes, it is, because it has a triangle with a dot inside it. So draw a circle around "yes" in row C under 1. Now is picture number two a Dax? No, it is not, because the dot is outside the triangle. So you draw a circlexaround the "no" in row C under 2. Now you go ahead and do the rest. (Pause. Test replaced by fourth Dax)

Let us look at the fourth example. Again, we see that a Dax is a triangle with a dot inside it. Alright now, you do the fourth puzzle. (Dax replaced by test) Now since there is a time limit on our puzzles, we are going to give you just the amount of time you will have for the other problems. You will then have an idea of how fast you must work. (25 second interval) Now I am going to tell you

the answers to this fourth puzzle, and you see if you had them right. Number one is a Dax; number two is not a Dax; etc.

Alright now. The Dax was only a practice puzzle. The puzzles you try from now on will be counted. You will work at each one just as you did with the Dax.

There is just one more rule in solving these puzzles: Once a new picture has been shown do not go back and make any changes in answers that you have already made. If you do, those answers will be counted wrong. For example, in this next puzzle, let us say you have been shown the first picture of a Mef and that you have already done the first puzzle..... that is, you have finished row A. When the second picture of a Mef is shown you must not go back and make any changes in that first row.

Now, everyone ready.

B. Instructions Issued Subjects In Experiment II.

Today your teacher has suggested that you help us work out some picture-puzzles. We think you will enjoy doing these puzzles. You have never seen them before. To make it more interesting we shall score your results. Here on the black-board you see the first part of your ahswer sheet: Fill in your name, whether a boy or girl, your age, birthday, school, and the name of your teacher. Pay no attention to the other blanks.

First on the screen we are going to show you a picture of a thing called a Dax (Spell). You will study this picture for a few moments to discover the idea of what a Dax is. Then you will be shown the puzzle made up of 10 pictures. You are to tell which of these 10 pictures contain the idea of what a Dax is, and which do not.

To make this clear, let us look at the first example. (Dax flashed on screen) Here is a picture of a Dax. Now study it and see if you can decide what a Dax is. (Pause. Then Dax replaced by test) Now look at the puzzle. Look at number one. Do you think it contains the idea of a Dax? If you do then draw a circle around "yes" in row A under 1. If you think it does not contain the idea of a Dax, circle the "no" in row A under 1. (Illustrating) Then look at picture number two. Do you think it contains the idea of a Dax? If you do then draw a circle around "yes" in row A under 2. If you think it does not contain the idea of a Dax, then circle the "no" in row A under 2. (Illustrating) Now you do the same for the other pictures. Draw the circles around "yes" or "no" <u>in row A</u>, because this is the first puzzle. Pay no attention to these columns on the right. <u>Be sure that you put your answers under "DAX" on your sheet</u>. (Pause) Raise your hand when you have finished. (Test replaced by non-Dax)

Now let us look at the second example on the screen. Here we have something that is not a Dax. Now you must

remember what the Dax was like, and see how this example is different from it. You remember that the other example had a triangle and a dot. But this one has only a triangle. Now let us look at the second puzzle. (Non-Dax replaced by test) Look at picture number one. Do you think it contains the idea of a Dax? If you do, then circle the "yes" in row B under 1, because this is the second puzzle. If you think picture number one does not contain the idea of a Dax, then circle the "no". (Illustration) Now look at picture number two. Do you think it is a Dax? If you do, then circle the "yes" in row B under 2. If you think it is not a Dax, circle the "no". (Illustrating) Now you do the rest of the puzzle. (Pause. Test replaced by second Dax)

Let us look at the third example on the screen. Now this one is a Dax. You remember the first example of a Dax that you saw had a triangle and a dot. So has this one. Remember that the shape of the triangle in the other Dax was not the same as this one, so that the shape of the triangle does not matter. You remember also that the dot in the other Dax was in a different position, so that the position of the dot does not matter. Now study this Dax closely. Do you think that a Dax is a triangle and a dot? Well, I am going to tell you what a Dax really is. A Dax is a triangle with a dot inside it. You remember that in the first Dax the dot was also inside the triangle. But

the second example had no dot, so that it was not a Dax. Now you do the third puzzle. (Pause. Dax replaced by test) Do you think picture number one is a Dax? Yes, it is, because it has a triangle with a dot inside it. So draw a circle around "yes" in row C under 1. Now is picture number two a Dax? No, it is not, because the dot is outside the triangle. So you draw a circle around the "no" in row C under 2. Now you go ahead and do the rest. (Pause. Test replaced by non-Dax)

Let us look at the fourth example. Now this is not a Dax, because the dot is outside the triangle. Alright now, you do the fourth puzzle. (Non-Dax replaced by test) Now since there is a time limit on our puzzles, we are going to give you just the amount of time you will have for the other problems. You will then have an idea of how fast you must work. (25 second interval) Now I am going to tell you the answers to this fourth puzzle, and you see if you had them right. Number one is a Dax; number two is not a Dax; etc.

Alright now. The Dax was only a practice puzzle. The puzzles you try from now on will be counted. You will work at each one just as you did with the Dax.

There is just one more rule in solving these puzzles: Once a new picture has been shown do not go back and make any changes in answers that you have already made. If you do, those answers will be counted wrong. For ex-

ample, in this next puzzle, let us say you have been shown the first picture of a Mef and that you have already done the first puzzle..... that is, you have finished row A. When the second picture of a Mef is shown you must not go back and make any changes in that first row.

Now, everyone ready.

C. Instructions Issued Subjects In Experiment III.

Today your teacher has suggested that you help us work out some picture-puzzles. We think you will enjoy doing these puzzles. You have never seen them before. To make it more interesting we shall score your results. Here on the black-board you see the first part of your answer sheet: Fill in your name, whether a boy or girl, your age, birthday, school, and the name of your teacher. Pay no attention to the other blanks.

First on the screen we are going to show you a picture of a thing called a Dax (Spell). You will study this picture for a few moments to discover the idea of what a Dax is. Then you will be shown the puzzle made up of 10 pictures. You are to tell which of these 10 pictures contain the idea of what a Dax is, and which do not.

To make this clear, let us look at the first example. (Dax on) Here is a picture of a Dax. Now study it and see if you decide what a Dax is. (Pause. Then test also flashed on) Now look at the puzzle. Look at

Number One. Do you think it contains the idea of a Dax? If you do then draw a circle around "yes" in row A under "1". If you think it does not contain the idea of a Dax, circle the "no" in row A under "1". (Illustrating) Then look at picture Number 2. Do you think it contains the idea of a Dax? If you do then draw a circle around "yes" in row A under "2". If you think it does not contain the idea of a Dax, then circle the "no" in row A under "2". Now you do the same for the other pictures. Draw the circles around "yes" or "no" in row A, because this is the first puzzle. Pay no attention to these columns on the right. Be sure that you put your answers under "Dax" on your sheet. (Pause) Raise your hand when you have finished. (Test off. First positive Dax supplemented by a negative Dax)

Let us look at the two examples on the screen. The top example is the Dax that you just studies. Now, below it is an example of something that is not a Dax. You will notice that the Dax (indicating) has a triangle and a dot, while the example below has only a triangle. We might guess, then, that a Dax is a dot and a triangle. Now let us look at the second puzzle. (Test flashed on) Look at picture Number One. Do you think it contains the idea of a Dax? If you do, then circle the "yes" in row B under "1", because this is the second puzzle. If you think picture Number 1 does not contain the idea of a Dax, then

circle the "no". (Illustrating) Now look at picture Number 2. Do you think it is a Dax? If you do, then circle the "yes" in row B under "2". If you think it is not a Dax, circle the "no". Now you do the rest of the test. (Pause. Test off, and two preceding examples supplemented by a third example).

Let us look at the three examples on the screen. You have already studied the first two (indicating). Below them is another example of a Dax. You will notice that the first Dax (indicating) has a triangle and a dot. So has this one. Notice also that the shape of the triangle in the first Dax is not the same as this one, so that the shape of the triangle does not matter. You can see, too, that the dot in the first Dax is in a different position, so that the position of the dot does not matter. Now study this Dax closely. (referring to third example). Do you think that a Dax is a triangle and a dot? Well, I am going to tell you what a Dax <u>really</u> is. A Dax is a triangle <u>with a dot inside it</u>. Note that in the first Dax the dot is also inside the triangle. But the second example has no dot, so that it is not a Dax. Now you do the third puzzle. (Test flashed on). Do you think picture Number One is a Dax? Yes, it is, because it has a triangle with a dot inside it. So draw a circle around "yes" in row C under "1". Now is picture Number 2 a Dax? No, it is not, because the dot is outside the triangle. So you

draw a circle around the "no" in row C under "2". Now you go ahead and do the rest. (Pause. Test off, and a fourth example added)

Now let us study the four examples on the screen. You are familiar with the first three. But if you look at the last example you can see that it is not a Dax, because the dot is outside the triangle. Alright, now, you do the fourth puzzle. (Test flashed on) Now since there is a time limit on our puzzles, we are going to give you just the amount of time you will have for the other problems. You will then have an idea of how fast you must work. (25 second interval) Now I am going to tell you the answers to this fourth puzzle, and you see if you had them right. Number 1 is a Dax. Number 2 is not a Dax. Etc., etc.

Alright now. The Dax was only a practice puzzle. The puzzles you try from now on will be counted. You will work at each one just as you did with the Dax.

There is just one more rule in solving these puzzles: Once a new picture has been shown do not go back and make any changes in answers that you have already made. If you do, those answers will be counted wrong. For example, in this next puzzle, let us say you have been shown the first picture, of a Mef and that you have already done the first puzzle..... that is, you have finished row A. When the second picture of a Mef is shown you must not go back and make any changes in that first row.

Alright now. Everyone ready.

APPENDIX II.

TABLE A. PERFORMANCE OF HIGH IQ GROUPS ON TOTAL OF TESTS A, B, C, AND D. MEAN SCORES, STANDARD DEVIATIONS, AND STANDARD ERRORS.

	GROUP I		GROUP II		GROUP III	
	Boys	Girls	Boys	Girls	Boys	Girls
A.M.	219.50 (6.39)	224.83 (7.33)	220.83 (8.56)	220.83 (9.36)	226.83 (10.96)	241.50 (4.12)
S.D.	23.90 (4.52)	27.40 (5.48)	32.00 (6.05)	35.00 (6.62)	41.00 (7.75)	15.40 (2.91)

TABLE B. PERFORMANCE OF LOW IQ GROUPS ON TOTAL OF TESTS A, B, C, AND D. MEAN SCORES, STANDARD DEVIATIONS, AND STANDARD ERRORS.

	GROUP I		GROUP II		GROUP III	
	Boys	Girls	Boys	Girls	Boys	Girls
A.M.	209.50 (4.57)	198.17 (5.56)	180.83 (9.25)	181.50 (5.75)	180.17 (9.52)	220.83 (7.73)
S.D.	17.10 (3.23)	20.80 (3.93)	34.60 (6.54)	21.50 (4.06)	35.60 (6.73)	28.90 (5.46)

TABLE C. ORDER OF ITEM-DIFFICULTY FOR HIGH IQ GROUPS IN TERMS OF PERCENTAGE OF MAXIMUM POSSIBLE SCORE.

GROUP I				GROUP II				GROUP III			
Boys		Girls		Boys		Girls		Boys		Girls	
Item	%	Item	%	Item	%	Item	%	Item	%	Item	%
Zum	85	Zum	92	Zum	82	Zum	82	Zum	85	Zum	93
Vec	84	Vec	83	Vec	80	Vec	78	Vec	81	Vec	88
Mef	72	Mef	75	Zif	72	Mef	72	Zif	78	Zif	79
Zif	69	Zif	70	Wez	71	Wez	70	Wez	73	Mef	79
Wez	64	Mib	66	Mef	70	Zif	68	Mef	73	Wez	76
Tov	59	Wez	64	Mib	60	Mib	65	Pog	63	Pog	66
Mib	59	Tov	60	Pog	59	Pog	61	Mib	62	Mib	63
Pog	56	Pog	53	Tov	52	Tov	55	Tov	52	Tov	55

TABLE D. ORDER OF ITEM-DIFFICULTY FOR LOW IQ GROUPS IN TERMS OF PERCENTAGE OF MAXIMUM POSSIBLE SCORE.

GROUPS I				GROUP II				GROUP III			
Boys		Girls		Boys		Girls		Boys		Girls	
Item	%	Item	%	Item	%	Item	%	Item	%	Item	%
Zum	83	Zum	74	Zum	66	Zum	69	Zum	70	Zum	86
Vec	83	Mef	67	Wez	59	Vec	61	Wez	60	Vec	79
Mef	70	Vec	63	Zif	58	Wez	60	Zif	58	Mef	71
Wez	59	Tov	60	Vec	57	Zif	56	Mef	57	Zif	68
Zif	58	Mib	59	Tov	53	Pog	54	Vec	56	Wez	66
Tov	57	Wez	57	Mef	53	Tov	51	Pog	54	Pog	62
Mib	57	Pog	55	Mib	52	Mef	50	Tov	49	Mib	60
Pog	56	Zif	54	Pog	52	Mib	50	Mib	46	Tov	56

TABLE E. COMPARISON OF NUMBER OF PERFECT SCORES IN TEST A WITH NUMBER OF PERFECT SCORES CONTINUING THROUGHOUT TESTS A, B, C, AND D. (SUB-GROUPS ONLY).

	GROUP I						GROUP II						GROUP III					
	Boys			Girls			Boys			Girls			Boys			Girls		
	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L
PERFECT SCORES IN TEST A.	17	8	7	17	10	9	12	11	12	16	8	3	24	14	7	20	12	13
PERFECT SCORES THROUGHOUT ALL FOUR TESTS	11	4	4	9	3	0	10	6	8	11	4	1	18	8	0	18	5	9

TABLE F. CRITICAL RATIOS OF THE DIFFERENCES IN VARIABILITY BETWEEN A, B, C, and D SCORES WITHIN EACH GROUP.

	GROUP I		GROUP II		GROUP III	
	Boys	Girls	Boys	Girls	Boys	Girls
A & B	.43	.70	3.31	3.84	2.00	1.54
B & C	1.72	.00	2.51	2.25	1.03	2.30
C & D	.07	1.13	3.48	2.92	2.63	1.04
A & C	1.14	.69	.82	1.73	1.42	3.48
B & D	1.74	1.20	1.67	1.43	2.23	1.69

TABLE G. HIGH GROUP PERFORMANCE ON TOTAL OF EACH OF TESTS A, B, C, AND D. MEAN SCORES, STANDARD DEVIATIONS, AND STANDARD ERRORS.

		GROUP I		GROUP II		GROUP III	
		Boys	Girls	Boys	Girls	Boys	Girls
TEST A	A.M.	51.43 (1.64)	52.50 (1.79)	53.83 (1.59)	55.17 (1.94)	56.77 (1.49)	55.97 (1.60)
	S.D.	6.12 (1.16)	6.68 (1.26)	5.96 (1.13)	7.24 (1.37)	5.56 (1.05)	6.00 (1.13)
TEST B	A.M.	55.17 (1.59)	55.70 (2.11)	52.23 (2.86)	53.83 (2.55)	55.17 (3.09)	58.90 (1.18)
	S.D.	5.96 (1.13)	7.88 (1.49)	10.68 (2.02)	9.52 (1.80)	11.56 (2.19)	4.40 (.83)
TEST C	A.M.	56.77 (1.94)	58.10 (2.24)	57.57 (2.19)	57.03 (2.59)	58.90 (3.07)	62.10 (1.09)
	S.D.	7.24 (1.37)	8.36 (1.58)	8.20 (1.55)	9.68 (1.83)	11.48 (2.17)	4.08 (.77)
TEST D	A.M.	56.23 (2.01)	57.30 (2.18)	57.30 (2.95)	55.17 (3.36)	56.50 (4.35)	62.37 (.94)
	S.D.	7.52 (1.42)	8.16 (1.54)	11.04 (2.09)	12.56 (2.37)	16.28 (3.08)	3.52 (.67)

TABLE H. LOW GROUP PERFORMANCE ON TOTAL OF EACH OF TESTS A, B, C, AND D. MEAN SCORES, STANDARD DEVIATIONS, AND STANDARD ERRORS.

		GROUP I		GROUP II		GROUP III	
		Boys	Girls	Boys	Girls	Boys	Girls
TEST A	A.M.	51.70 (1.80)	48.23 (1.89)	46.10 (1.95)	47.17 (1.39)	43.70 (2.72)	53.03 (2.02)
	S.D.	6.72 (1.27)	7.08 (1.34)	7.28 (1.38)	5.20 (.98)	10.16 (1.92)	7.56 (1.43)
TEST B	A.M.	52.77 (1.33)	49.03 (1.35)	42.37 (2.95)	37.57 (2.09)	42.90 (2.24)	53.83 (2.33)
	S.D.	4.96 (.94)	5.04 (.95)	11.02 (2.08)	7.80 (1.47)	8.36 (1.58)	8.72 (1.65)
TEST C	A.M.	51.97 (1.51)	49.30 (1.47)	48.50 (2.11)	52.77 (2.09)	47.43 (2.48)	58.37 (1.74)
	S.D.	5.64 (1.07)	5.48 (1.04)	7.88 (1.49)	7.80 (1.47)	9.28 (1.75)	6.52 (1.23)
TEST D	A.M.	52.50 (1.61)	49.30 (1.83)	43.97 (3.47)	43.43 (2.51)	46.10 (3.30)	56.23 (2.91)
	S.D.	6.04 (1.14)	6.84 (1.29)	12.98 (2.45)	9.40 (1.78)	12.36 (2.34)	10.88 (2.06)

TABLE I. CRITICAL RATIOS OF THE DIFFERENCES BETWEEN MEAN SCORES WITHIN HIGH GROUPS. (BRACKETED LETTERS DESIGNATE HIGHEST SCORES).

	GROUP I		GROUP II		GROUP III	
	Boys	Girls	Boys	Girls	Boys	Girls
A & B	3.20(B)	1.83(B)	.86(A)	.68(A)	.66(A)	2.64(B)
B & C	1.72(C)	2.73(C)	2.37(C)	3.95(C)	1.59(C)	5.08(C)
A & C	4.49(C)	2.84(C)	2.09(C)	.87(C)	1.01(C)	4.41(C)
B & D	1.13(D)	2.13(D)	2.41(D)	1.09(D)	.91(D)	4.34(D)

TABLE J. CRITICAL RATIOS OF THE DIFFERENCES BETWEEN MEAN SCORES WITHIN LOW GROUPS. (BRACKETED LETTERS DESIGNATE HIGHEST SCORES).

	GROUP I		GROUP II		GROUP III	
	Boys	Girls	Boys	Girls	Boys	Girls
A & B	.47(B)	.63(B)	2.10(A)	4.28(A)	.03(A)	.04(B)
B & C	.61(B)	.24(C)	3.56(C)	5.78(C)	2.38(C)	2.27(C)
A & C	.17(C)	.75(C)	1.32(C)	2.83(C)	2.02(C)	4.60(C)
B & D	.18(B)	.23(D)	.82(D)	3.64(D)	1.99(D)	2.12(D)