INFORMATION TO USERS

This was produced from a copy of a document sent to us for microfilming. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help you understand markings or notations which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure you of complete continuity.

2. When an image on the film is obliterated with a round black mark it is an indication that the film inspector noticed either blurred copy because of movement during exposure, or duplicate copy. Unless we meant to delete copyrighted materials that should not have been filmed, you will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed the photographer has followed a definite method in "sectioning" the material. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For any illustrations that cannot be reproduced satisfactorily by xerography, photographic prints can be purchased at additional cost and tipped into your xerographic copy. Requests can be made to our Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases we have filmed the best available copy.

University Microfilms International
300 N. ZEEB ROAD, ANN ARBOR, MI 48106
18 BEDFORD ROW, LONDON WC1R 4JF, ENGLAND
CONTEXTUAL DETAILS AND MATHEMATICAL
PROBLEM DIFFICULTY

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate
School of The Ohio State University

By

Charles William Zimmerman, B.S., M.A.

* * * * *

The Ohio State University

1979

Reading Committee:
Professor Richard J. Shumway
Professor Harold C. Trimble
Professor John W. Riner

Approved by
Advisor
College of Education
ACKNOWLEDGMENTS

I would like to thank the teachers and students of the classes who participated in the two experiments.

It is impossible to acknowledge adequately my debt to the members of my reading committee, Professors Richard Shumway, John Riner and Harold Trimble. Richard Shumway, my major advisor and good friend, has provided invaluable advice during the completion of this dissertation. He has helped me to realize the importance of educational research and has been a consistent example of professional competence. Harold Trimble has helped me to gain insights into many issues. I am grateful for our many discussions. To John Riner I owe many things. He gave me the opportunity to gain many experiences teaching with the Department of Mathematics and it was an immense pleasure working with him.

I would like to thank my family and friends whose encouragement made the work possible. To Amy thanks for helping to point out some factors that increase problem difficulty from the student's point of view.

Finally, thanks to Flora Cooney who typed the manuscript. Her work was excellent and I appreciate the care she took with the task.
VITA

April 2, 1947........Born - Huron, Ohio

1968........B.S. in Mathematics Education
The Ohio State University, Columbus, Ohio

1969 - 1972........Secondary School Mathematics Instructor, Upper Arlington, Ohio and Berea, Ohio

1974........M.A. in Mathematics, The Ohio State University, Columbus, Ohio

1972 - 1978........Teaching Associate, Lecturer, Department of Mathematics, The Ohio State University, Columbus, Ohio

FIELDS OF STUDY

Major Field: Mathematics Education

Studies in Mathematics Education. Professors Richard Shumway and Harold Trimble

Studies in Mathematics. Professor John Riner

iii
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>ii</td>
</tr>
<tr>
<td>VITA</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>II. BACKGROUND OF THE PROBLEM</td>
<td>19</td>
</tr>
<tr>
<td>III. METHODS AND PROCEDURES</td>
<td>38</td>
</tr>
<tr>
<td>IV. DATA ANALYSIS</td>
<td>73</td>
</tr>
<tr>
<td>V. CONCLUSIONS</td>
<td>97</td>
</tr>
<tr>
<td>APPENDICES</td>
<td></td>
</tr>
<tr>
<td>A. Problems Solved During Treatments for Experiment I</td>
<td>116</td>
</tr>
<tr>
<td>B. Test Items for Experiment I</td>
<td>129</td>
</tr>
<tr>
<td>C. Problems Solved During Treatments for Experiment II</td>
<td>134</td>
</tr>
<tr>
<td>D. Test Items for Experiment II</td>
<td>153</td>
</tr>
<tr>
<td>E. Sample Interaction Between Researcher and Subjects</td>
<td>162</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>166</td>
</tr>
<tr>
<td>Table</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>1. Means and Standard Deviations for Small Group Averages on Test AB and Test BA</td>
<td>77</td>
</tr>
<tr>
<td>2. Multivariate and Univariate Analysis of Variance Scores for Variables Test AB, Test, BA, Using Wilks Lambda Criterion</td>
<td>77</td>
</tr>
<tr>
<td>3. Small Group Averages for Test AB and Test BA</td>
<td>80</td>
</tr>
<tr>
<td>4. Means and Standard Deviations for Errors of Misclassification</td>
<td>80</td>
</tr>
<tr>
<td>5. Multivariate and Univariate Analysis of Variance Errors of Misclassification for Variables Test AB, Test BA Using Wilks Lambda Criterion</td>
<td>81</td>
</tr>
<tr>
<td>6. Small Group Averages for Errors of Misclassification</td>
<td>83</td>
</tr>
<tr>
<td>7. Small Group Subtest Means</td>
<td>87</td>
</tr>
<tr>
<td>8. Treatment Means and Standard Deviations for the Four Subtests</td>
<td>88</td>
</tr>
<tr>
<td>9. Multivariate and Univariate Analysis of Variance Scores for the Four Subtests Using Wilks Lambda Criterion</td>
<td>89</td>
</tr>
<tr>
<td>10. Means and Significant Comparisons for the Four Subtests</td>
<td>91</td>
</tr>
<tr>
<td>11. Treatment Means for High Ability Subjects</td>
<td>94</td>
</tr>
<tr>
<td>12. Treatment Means for Low Ability Subjects</td>
<td>94</td>
</tr>
<tr>
<td>13. Means and Significant Comparisons for Small Group Averages on Test AB and Test BA</td>
<td>99</td>
</tr>
</tbody>
</table>
Table Page

14. Means and Significant Comparisons for Errors of Misclassification.......................... 100

15. Treatment Means and Significant Comparisons for the Four Subtests......................... 104

16. Results When Grouped by the Number of Sessions Using Each Problem Form................ 105
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Design for Experiment I</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>Design for Experiment II</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>Order of the Data for Each Type-Form Combination</td>
<td>63</td>
</tr>
<tr>
<td>4</td>
<td>Design for Experiment I</td>
<td>74</td>
</tr>
<tr>
<td>5</td>
<td>Plots of Treatment Means for Subtest Scores</td>
<td>78</td>
</tr>
<tr>
<td>6</td>
<td>Plots of Treatment Means for Errors of Misclassification</td>
<td>82</td>
</tr>
<tr>
<td>7</td>
<td>Design for Experiment II</td>
<td>84</td>
</tr>
<tr>
<td>8</td>
<td>Plots of Treatment Means</td>
<td>89</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

Polya (1957) suggested in his classic on problem solving, *How to solve it*, that when devising a scheme to solve a problem it is useful to think of a "related" problem. Undoubtedly, many mathematics teachers make this suggestion to their students as a means of solving problems. But how do students react to this advice? What problems are thought of as "related" problems? Although it is not fair to say Polya recommends that students mimic solutions to previous problems, instruction may leave students believing that they should solve problems in this manner. Duhmas (1970) stated in an article entitled "How to teach verbal problems," that verbal problems were frequently categorized as "percent problems, mixture problems, work problems, and others." However, Duhmas noted that, "Classifying problems encourages students to look for relatively unimportant common characteristics. At best, most of the classifications serve only as temporary crutches for a few students."
Krutetskii (1976) determined that above average subjects are likely to classify problems according to structural similarities. However, below average subjects tend to classify problems according to relatively unimportant contextual details. Therefore, Krutetskii has predicted that when "good" students recall a related problem, they recall a problem which was solved using the same mathematics. But, "weak" students often recall a problem involving the same contextual details. It is also likely that weak students will not recall any related problems if no problems involving the same contextual details have been encountered.

This study investigated the effect of certain irrelevant properties including contextual details on the difficulty of arithmetic word problems. Two distinct but related experiments were conducted using disjoint samples. The use of disjoint samples was deemed necessary to assure that there were no interactions between the two experiments. The problem sets used for the instructional treatments differed on the irrelevant attributes and contextual details, but the mathematical content of each of these instructional treatments was identical. Instruction was given to small groups of approximately seven students.

In the first experiment, the possibility of negative transfer caused by associating contextual details and the
method of solution was studied. The second experiment investigated the extent to which subjects transferred knowledge to solve problems whose irrelevant properties were not matched to those encountered during instruction. The impact of familiarity with problem appearance was measured by observing student performance on problems whose contextual details and irrelevant attributes were either matched or not matched with the problems encountered during the instructional treatments.

NEED FOR THE STUDY

Kilpatrick (1969) concluded in his review of the problem solving research in mathematics that, "Few studies build on previous research; few studies have an explicit theoretical rational." A major weakness of the research in mathematics education is the lack of empirical studies that verify the predictions based on the theoretical models.

The development of the ability to solve previously unencountered problems is frequently a goal of mathematics instruction. Often mathematics students are capable of solving a problem in one context, and yet fail to solve an "equivalent" problem in a different context. Bourne and Dominowski (1972) agree that "solution familiarity, and the manner in which the problem solver is given the information are key factors in determining
problem difficulty." Bourne (1972) also noted, "little has been done to evaluate solution familiarities effect on success." Markle (1969, 1970) stressed that development of the ability to detect small differences and similarities between problems should be a function of instruction.

Merrill (1971, 1973) has developed a cognitive model which may help to explain why many students do not transfer the knowledge to solve previously unencountered problems. He defines four levels of cognitive behavior as follows: (1) emotional, (2) psychomotor, (3) memorizational, and (4) complex cognitive. The memorizational and complex cognitive levels may be helpful in understanding mathematics learning.

When a student responds to a question that is identical in appearance to one that he has previously encountered, he is required to display only a "discriminated recall" of the specific information. Thus the student is operating at the "memorizational" level. However, to solve an example which has a previously unencountered appearance, requires that the student "classify" the problem, this is "complex cognitive" behavior. Classification occurs when a student correctly identifies the class membership of a previously unencountered object, event, or symbol. Merrill stresses that no amount of practice with "discriminated recall" will develop the
ability for cognitive transfer.

One interpretation of this model is that solving a mathematics problem which is identical except for the numbers to a previously encountered problem, requires only "memorizational" behavior. But if a problem is stated in a form which the student has not encountered, then the problem is more difficult because "complex cognitive" behavior is required to solve the problem. A more detailed discussion of this model will be presented in Chapter II.

Clark (1971) concluded in a review of the psychological research literature that "The form of the instance apparently determines the number of non-critical properties that will be evoked by it. As this number decreases, ease of concept attainment increases." Clark concluded that the greatest transfer of prior knowledge occurs for:

(a) new instances of the same concept, and
(b) new but related concepts when a variety of original instances are learned and each of these original instances is completely learned.

Callentine (1955) and Adams (1954) have examined the effect of using multiple instances of the same concept during instruction. Other studies (Huffman, 1963; Bourne and Guy, 1968; Wittrock, 1973) have determined that the form of the problem used during instruction affects the problem difficulty for the criterion test items. These
studies are discussed in more detail in Chapter II.

Kilpatrick (1969) points out that very little research has examined how solution familiarity and problem appearance affect the difficulty of mathematics problems. Several Soviet studies (Yaroskchuk, 1969; Krutetskii, 1969, 1976) have examined the tendency to identify problems which they are solving as types which they have previously solved. Evans and Pike (1973) investigated the conjecture that special instruction can improve a subject's ability to solve problems with a specified format. They determined that students scored better on the items which employed the format used during the special instruction.

The order in which the data is presented in a problem appears to be a factor in determining problem difficulty. The results of Burns (1964), Williams (1965), and Rosenthal (1971) relevant to this question will be discussed more specifically in Chapter II.

The reasons why a familiar appearance in a story problem results in the problem being less difficult is unclear, but reading ability seems to be a factor. Call (1966), Minski (1975), Filano (1951), and Anderson (1977) are among the studies which provide support for the conjecture that reading ability and problem familiarity are interrelated. It is possible that if a student has previously encountered an instance of a problem, his experience with the contextual details enables him to
form a story schemata and therefore, a better understanding of certain relationships between the variables. This, in turn, translates into greater success on similar problems.

The studies which have been reported in this research have generally indicated that problem appearance, prior familiarity, order of data, and reading ability may be important in determining problem difficulty. Merrill's model may provide the theoretical framework which explains the importance of prior familiarity with a problem form. Markle (1970) has stressed the need to vary the irrelevant attributes when selecting instances for use during instruction. Markle (1970) and Tennyson (1972a, 1972b, 1973) have presented recommendations based on Merrill's model for deriving example items which will avoid student misconceptions. First, the irrelevant attributes should be varied for instances of the same problem type. Secondly, the irrelevant attributes should be matched for instances of different problem types. Thus, there are some specific suggestions on proper construction of instructional examples. However, almost no controlled research has directly examined how variability among the irrelevant properties effects a students ability to transfer knowledge to solve problems. Nor has research examined the effect of prior familiarity with a problems appearance on problem difficulty. The
The purpose of this research was to investigate these effects in a controlled setting.

OBJECTIVES

The purpose was to determine the significance of prior familiarity with specific irrelevant properties on the difficulty of selected word problems. In this study "familiarity" with a problem form means that the subject has previously solved a problem that used the same problem form. The treatments presented during the study differed on the form of the word problems used during instruction. Different problem forms were derived by systematically varying irrelevant properties. By measuring student performance on different problem forms it was hoped to estimate the importance of prior familiarity with the problem form. The effect of familiarity with problem appearance on the problem difficulty was estimated under two different situations.

The first experiment investigated the extent that the problem appearance became associated with the method of solution when the irrelevant properties were not systematically varied. Differences between the treatments were anticipated for two reasons. First, positive transfer could have resulted for problems which employed the same type-form combinations encountered during the treatments. Second, due to associations between the
problem form and the algorithm, negative transfer could have occurred for problems which were not matched to instances encountered during the instructional sessions.

The second experiment investigated the effect of prior familiarity with problem forms and the variability of the problem forms during instruction on a student's ability to transfer. In this case, negative transfer was unlikely because three problem types were given using each problem form. Therefore, differences between the scores for the different treatments were attributed to the problem forms encountered during instruction since only the problem forms had been varied.

SETTING

As noted earlier the study consisted of two separate experiments. The subjects for the first experiment were 85 fifth-grade students from two Columbus area elementary schools. The subjects used in the second experiment were 213 sixth-grade students from three Columbus area schools. In both experiments the mathematical content encountered during the instructional treatments was not closely related to the material covered in the students' curriculum. The instructional sessions were presented to groups of approximately seven students in a small room. Each session lasted for approximately thirty minutes. A black board was used to present and discuss
problems during the instructional sessions.

EXPERIMENT I

The first experiment was designed to document that the appearance of a problem is frequently associated with the algorithm used to solve the given problem. One conjecture was that when the problem type-form combinations encountered by a subject were not systematically varied, the subject would identify a specific computational algorithm with the problems appearance. Therefore, when the subject encounters a problem with an appearance similar to one previously encountered, he will recall the previous problem and apply the algorithm he applied to the recalled problem even if that algorithm is not appropriate for this new problem.

Two problem forms and two problem types were derived for the two instructional treatments in Experiment II. Thus a total of four problem type-form combinations were constructed. These combinations were denoted as Type I - Form A, Type I - Form B, Type II - Form A, and Type II - Form B. Problems were of the same type if they could be solved using the same algorithm. When problems were matched on the contextual details and other irrelevant attributes affecting problem appearance, the problems were said to have the same problem form.
Subjects who received Treatment AB encountered Type I - Form A and Type II - Form B problems during the instructional sessions. Subjects who received Treatment BA encountered Type I - Form B and Type II - Form A problems during the instructional sessions.

The population was partitioned into 12 small groups. Four small groups received a first instructional treatment; a second four small groups received a second instructional treatment; and four small groups served as control groups and received no instruction. Each group which received an instructional treatment participated in two instructional sessions. During each session, the subjects solved several word problems. The difference between the treatments was determined by the irrelevant properties present in the problem statements. The primary irrelevant properties which were varied included: the contextual details, the order of the data, type of units, and the use of certain key words. The two instructional treatments were denoted as Treatment AB and Treatment BA respectively. After the instructional treatments were finished, each subject completed a twelve item test which, for the purpose of analysis, was subdivided into two six item subtests. The items of one subtest matched the problem type-form combinations used during the instructional sessions for Treatment AB. This subtest was denoted as Test AB. The second subtest,
denoted as Test BA, consisted of items which matched the problem type-form combinations solved during Treatment BA.

The experimental design is indicated in Figure 1. A more detailed description of the treatments and procedures will be presented in Chapter III. A multivariate analysis of variance was employed to analyze the data. Appropriate univariate tests were applied to determine the nature of the differences between the treatments. Results are presented in Chapter IV.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Test AB</th>
<th>Test BA</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>BA</td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>CONTROL</td>
<td>X</td>
<td>Y</td>
</tr>
</tbody>
</table>

FIGURE 1
DESIGN FOR EXPERIMENT I

EXPERIMENT I

HYPOTHESES, ASSUMPTIONS, AND LIMITATIONS

The hypotheses were:

(1) There were differences between the scores on Test AB for the three treatments.

(2) There were differences between the scores on Test BA for the three treatments.
(3) There were differences between the number of "errors of misclassification" on Test AB for the three treatments.

(4) There were differences between the number of "errors of misclassification" on Test BA for the three treatments.

The assumptions were: (1) The appearance of word problems as perceived by the subjects was affected by contextual details and other irrelevant properties more than many other factors. (2) The irrelevant properties which were systematically varied to derive the different problem forms were among the most critical irrelevant properties. (3) It was unlikely that the subjects had encountered problems which were highly similar to the problems introduced in the treatments.

The limitations were as follows: (1) Only word problems which required several steps to solve were used as items. (2) The subjects were 85 fifth-grade students from two Columbus area elementaries. (3) Only a subset of the irrelevant properties were varied to derive the different problem forms. (4) The instructional treatments were presented to small groups of only seven students. (5) Only two problem types were derived for the treatments. (6) Only two problem forms were derived for the treatments.
EXPERIMENT II

Assuming significant differences were observed between the instructional treatments in Experiment I, it is of interest to investigate instructional strategies which might minimize the tendency to associate the problem form with the solution algorithm. One strategy which may have the best chance to produce transfer is to vary the problem forms used during the instructional sessions.

Experiment II was designed to answer two questions. First, does familiarity with a problem's appearance cause the problem to be less difficult than a mathematically equivalent problem which uses an appearance that is unencountered by the student? Second, when the problem forms are systematically varied during the instructional sessions does this result in better performance on equivalent problems with an unencountered problem appearance?

Three problem types and four problem forms were derived. Thus a total of 12 problem type-form combinations were constructed. The problem forms were denoted as Form A, Form B, Form C, and Form D. Three instructional treatments were derived for Experiment II. The difference between the instructional treatments was determined by the forms used to present the set of problems during each session. The three treatments were
denoted as Treatment AAA, Treatment BBB, and Treatment ABC. Subjects who received Treatment AAA solved problems which employed Form A on each of the three days. The subjects that received Treatment BBB solved problems which used Form B each of the three instructional sessions. The subjects which received Treatment ABC encountered a different problem form during each session. Form A problems were solved during the first session, Form B problems during the second session, and Form C problems during the third session.

The population was partitioned into 29 small groups with 7 members each. None of the subjects had participated in Experiment I. Each of the three instructional treatments was presented to seven small groups. In addition, eight small groups did not receive any instruction, but did participate in the testing. These students served as a control population. Each small group receiving a treatment participated in three instructional sessions. Six problems were solved during each instructional session. The same three problem types were presented during each of the instructional sessions. The difference between the three instructional treatments was determined by the irrelevant properties used to derive the different problem forms. The irrelevant properties which were varied to derive four different problem forms included certain contextual details, the
order of the data, the use of certain key words, and the units employed in the problem statement. A more detailed description of the item construction is presented in Chapter III.

The four problem forms derived for Experiment II were denoted as Form A, Form B, Form C, and Form D. Form D problems appeared on the test, yet were not encountered by any subjects during the instructional treatments. This form was very important since it was used to estimate the performance on a previously unencountered problem form. A 24 item test was completed by each subject. The test consisted of four six-item subtests. The subtests differed on the problem form employed to write the items. The four subtests were denoted as Test A, Test B, Test C, and Test D. The experimental design is described in Figure 2.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Test A</th>
<th>Test B</th>
<th>Test C</th>
<th>Test D</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>W</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
</tr>
<tr>
<td>BBB</td>
<td>W</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
</tr>
<tr>
<td>ABC</td>
<td>W</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
</tr>
<tr>
<td>CONTROL</td>
<td>W</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
</tr>
</tbody>
</table>

**FIGURE 2**

DESIGN FOR EXPERIMENT II
A multivariate analysis of variance was employed to analyze the data. Appropriate univariate tests were applied to determine the nature of the differences between the treatments. Results are presented in Chapter IV.

EXPERIMENT II

HYPOTHESES, ASSUMPTIONS, AND LIMITATIONS

The hypotheses were:

(1) There were differences between the scores on Test A for the three instructional treatments.
(2) There were differences between the scores on Test B for the three instructional treatments.
(3) There were differences between the scores on Test C for the three instructional treatments.
(4) There were differences between the scores on Test D for the three instructional treatments.

The assumptions were: (1) The appearance of word problems was affected by contextual details and other irrelevant properties to a greater extent than many other problems. (2) The irrelevant properties which were varied to derive the different problem forms were among the most crucial irrelevant properties. (3) It was unlikely that the subjects had encountered problems which were highly similar to problems presented during the treatments.
The limitations were as follows: (1) Only word problems which required several steps to solve were used as items. (2) Only three problem types appeared in the treatments. (3) The subjects were 213 sixth-grade students from three Columbus area public schools. (4) Only a subset of the irrelevant properties were varied to derive the different problem forms. (5) Only four problem forms were used for this study. (6) The instructional treatments were presented to small groups of only seven subjects.
CHAPTER II

BACKGROUND OF THE PROBLEM

Few studies have investigated prior familiarity with contextual details as a factor in problem difficulty; but many studies have investigated related questions. Five areas in the literature which were related to this study will be summarized. First, one area consists of studies that investigated the students perceptions of problem similarity. A second area consists of studies that recommend ways to construct items designed to produce transfer. A third area consists of studies that have investigated factors which help predict problem difficulty. Fourth, there were studies which considered the subject's familiarity with specific problems. However, in this category the results were particularly difficult to interpret because "familiarity" had several different definitions. Finally, several studies on reading ability have implications for this study since the problems used were arithmetic word problems.
PERCEPTIONS OF SIMILARITY

Silver (1978) in his article "Student perceptions of relatedness among mathematical word problems" identified four fundamental problem similarity dimensions. These dimensions were: (a) the mathematical structure, (b) contextual details, (c) question form, and (d) pseudostructure. In Silver's study, the subjects were asked to sort problems into groups of similar problems. Each of the four dimensions were identifiable during the sorting task. The pseudostructure dimension referred "to associations based on the mutual presence of a measurable quantity, such as age, weight, or time."

That is, two problems would be related along the pseudostructure dimension if they contained the same units such as minutes. Subjects were asked to sort problems into sets of similar problems both before and after they were asked to solve the problems. Silver observed an increased awareness of the method of solution as a dimension in student perceptions on the post-solution sort.

Chartoff (1976) also investigated students' perceptions of mathematical problem similarity. He investigated students ranging in age from seventh grade to
college freshman. The subjects rated pairs of verbal problems on a scale from extremely similar to extremely dissimilar. The results indicated four dimensions for problem similarity almost identical to Silver's dimensions. For Chartoff, the dimensions were: (a) how the problems were solved, (b) the contextual setting, (c) comparisons with a generic problem of the same type, and (d) the question posed by the problem. In both studies, the contextual details were identified as the most significant dimension of problem similarity not affecting the method of solution.

Krutetskii (1969, 1976) has identified differences between high ability and low ability students in their perceptions of problem similarity. He reported that certain abilities are present in capable students, but absent in less capable students. In particular, capable students are able to distinguish relevant information from the contextual details in a problem statement. Capable students also tend to recall the mathematical structure of the problem, yet less capable students tend to remember details about the story related in the problem and not the problem structure. As Bright (1977) has indicated the capable students referred to by Krutetskii represent a very small portion of the student population. Perhaps less than five percent of the students in American schools could be classified as
"capable" students. But the results of Krutetskii's work has determined that how a student perceives contextual details is related to how successfully he solves the problems.

Brownell (1931) investigated the familiarity of problem appearance as a factor in determining problem difficulty. He found that the students' familiarity with the units and contextual details involved in the problem was a factor. Familiarity may be another dimension along which students perceive problems as similar. His definition of familiarity did not consider the students previous mathematical instruction, but rather estimated how familiar the contextual details might be "to the student," using a continuous scale from very familiar to very unfamiliar. The results indicated that the more familiar the units used in the problem statements were to the student, the better the problem was understood, which translated into better performance. However, these results were difficult to interpret due to the subjectiveness of his definitions and to certain ambiguities about the methods and procedures.

COGNITIVE DEVELOPMENT MERRILL'S MODEL

Merrill's model for cognitive development has provided the theoretical foundation for this research. Merrill (1971) defined four levels of cognitive behavior.
These were: emotional, psychomotor, memorizational, and complex cognitive. Two of these levels, memorizational and complex cognitive, are relevant when interpreting mathematical problem solving. Merrill has defined memorizational and complex cognitive behavior as indicated below.

Memorizational behavior occurs when a student immediately reproduces or recognizes without prompting, a specific symbolic response when presented with a specific stimulus situation. Complex cognitive behavior occurs when the student makes an appropriate response to a previously unencountered instance of some class of stimulus objects, events or situations.

Tennyson (1971), when comparing the cognitive models of Bloom, Gagne, and Merrill, concluded that Merrill's model more carefully considered the cue set used to present specific problems and the cue sets previously encountered by the subject. The occurrence of specific cue sets during instruction is particularly important in mathematics. One of the most frequent and widespread practices among mathematics teachers is to construct test items by choosing examples or homework exercises previously solved and simply changing the numbers in the problem statements. It is not likely that this practice will ever be eliminated; but according to Merrill's cognitive model, problems which are constructed by simply changing the numbers in previously encountered problems requires the student to operate only at the
memorizational level. Aiken (1973) also recognizes the importance of using examples which were previously unencountered by the subjects. He defined creativity in mathematics as the ability to "combine ideas in a new way."

Markle (1969), Tennyson (1972a, 1973), and Merrill (1971) have stressed the need to vary "irrelevant attributes" when constructing sets of instances for instruction. The research has usually focused on concept learning; but Merrill's Model for cognitive levels can also help predict problem solving behaviors. One of Merrill's subcategories under "memorizational" behavior is "multiple discrimination." "Problem solving" is a subcategory of complex cognitive behavior. Therefore, when a student encounters a problem which differs from previously solved problems only by the numbers in the problem statement he is required to display "multiple discrimination" or "discriminated recall" and operate only at the memorizational level. But when irrelevant properties which affect the cue set are also changed, "problem solving," a complex cognitive behavior, is required to solve the problem.

ITEM CONSTRUCTION AND TYPES OF ERRORS

Markle and Tiemann (1970) concluded, "When we expect a student to generalize to a completely new instance of a concept as a result of instruction, the new instance
cannot be a trick question of his ability to generalize. Similarly, in testing for discrimination, the ability to see small differences should be a function of the instruction that has taught the student how, not a function of his position as a good student." Tennyson and Boutwell (1974) and Tennyson, Woolley and Merrill (1972) have claimed that sets of items must be divergent and matched to avoid specific types of errors. For two problems of the same problem type to be divergent, the irrelevant attributes must be systematically varied. For two problems of different problem types to be matched, they must agree on many of the irrelevant attributes but differ on some of the relevant attributes. Although most of the work of Merrill, Tennyson, and Markle referred to concept learning, only minor changes in their definitions were needed to determine how their method of item construction could be applied to derive problem sets for mathematics.

Tennyson (1973) classified three types of errors that students commonly make when they classify instances of a concept. These errors are: (1) undergeneralization, (2) overgeneralization, and (3) misconception. A subject who undergeneralizes a concept, makes errors by rejecting certain examples of a concept as non-examples. The subject who overgeneralizes a concept, makes errors by accepting some non-examples of a concept as examples of the concept. The subject who displays a misconception
identifies some irrelevant attribute as a relevant attribute. On the basis of this irrelevant attribute he both rejects examples, and accepts non-examples of the concept. According to Tennyson, to produce a correct classification of instances of a concept, instruction must present a set of problems which are divergent and matched. Tennyson conjectured that when a set of instances of a concept is convergent and not matched, misconceptions will result for many students.

Definitions:

Convergent - Two examplars are convergent when the irrelevant attributes are as similar as possible.

Divergent - Two examplars are divergent when the irrelevant attributes are as dissimilar as possible.

Matched - A matched relationship between examplars and non-examplars occurs when the irrelevant attributes are as similar as possible.

Not Matched - An examplar and a non-examplar are not matched when the irrelevant attributes are as different as possible.

These definitions can be easily modified to apply to problem solving in mathematics.

Convergent - Two instances of the same problem type are convergent when the irrelevant attributes
are as similar as possible.

Divergent - Two instances of the same problem type are divergent when the irrelevant attributes are as dissimilar as possible.

Matched - A matched relationship exists between instances of two different problem types when the irrelevant attributes are as similar as possible.

Not Matched - Instances of two different problem types are not matched when the irrelevant attributes are as dissimilar as possible.

With these modifications, Tennyson would suggest that a proper set of exercises designed to teach students to solve problems would be divergent and matched. If a set of exercises consists of problems that are convergent and not matched, then misconceptions concerning the method of solution will result.

PROBLEM DIFFICULTY

Current research has determined several factors which may have an effect on problem difficulty. Rosenthal (1971) found two factors that affected student performance on arithmetic word problems. The factors were as follows: (a) the form of the problem, and (b) the sequence used to present the problem information. Aiken (1971) reported high correlations between problem solving and
reading ability. Nesher (1976) investigated four determinates of difficulty in verbal arithmetic problems. These determinates were the number of steps, the amount of extraneous information, the presence of verbal cue's, and the kind of story described in the question. He found that each of these determinates had a significant effect except for the kind of story described in the problem statement. This might appear to contradict the hypotheses of this study. However, Nesher did not investigate how these factors might have been influenced by previous instruction.

Williams (1965) did not find that the position of the question had an effect on the problem difficulty. Levin (1963), Roe (1962), and Niedermeyer (1969) did not find differences attributable to the order of the items presented during the instructional sessions. Burns and Yonally (1964) found that the order of the data in the problem statement did affect the difficulty. They determined that when the data was presented in the same order that it would be used in the solution that the difficulty was reduced.

When irrelevant information occurs in a problem statement there appears to be a dramatic increase in the difficulty. Overstreet (1969) and Johnson (1961) determined that the greater the frequency of irrelevant information in a problem statement, the more difficult the
problem. Bourne (1968) and Battig (1961) determined an almost linear relationship existed between the number of irrelevant dimensions in a problem and the difficulty of that problem.

FAMILIARITY WITH PROBLEM APPEARANCE

Few studies have looked at the question, "What is the effect on student performances of previously solving a problem with the same appearance?" Perhaps the study which has looked most directly at this question was Evans and Pike (1973). Evans and Pike investigated the conjecture that special instruction would result in improved performance on items with a specific format. They selected three different item formats which occurred on portions of the SAT Mathematics Placement Test. Of the three treatments, one treatment was specifically designed to help students score well on each of the item formats. The results indicated that each of the three treatment groups scored higher than the other groups on those items which matched the format of problems that had occurred in their treatment.

Adams (1954) found that multiple problem training increased the student's ability to solve problems previously unencountered during instruction. However, Callentine (1955) reported that a student trained in applying only one form of a concept, attained greater
proficiency with that form of the concept, yet was unable to transfer this knowledge to other forms of the same concept. Due to weaknesses in the design and reporting of these two studies, the implications are difficult to interpret.

Some studies have investigated the relationship between the problem forms used during instruction and the problem difficulty of criterion test items. Bourne and Guy (1968), when investigating the role of positive and negative instances, determined that subjects perform best when the information was presented in the same form as previously encountered instances. Huffman (1963) found that previous training in solving problems similar to the given problems improved student performance. However, experience with only simple problems seemed to be a hindrance in solving more difficult problems. Wittrock (1973) found that when students were accustomed to processing information in one way, they encountered difficulty when they were asked to process the information in a different manner. He concluded that when a student received the same information in two different forms, there was a tendency to treat the problems as two separate and unrelated problems.

Soviet studies have identified a tendency of students to categorize problems which they are solving as types which they have previously solved. Yaroskchuck (1969)
determined that when a student is able to correctly "typify" a problem by its mathematical structure, this student usually solved the problem correctly. However, when the student was unable to correctly "typify" the problem, he was rarely able to solve the problem without outside help. Krutetskii (1969, 1976), through the use of case studies, determined that the student's ability to generalize solutions of a single problem type was a factor which differentiates capable and less capable students. Krutetskii concluded that the average pupil learns to generalize by solving many problems of a single type with varying data. A less capable pupil will have extreme difficulty in solving problems of a single type, even with the experimenter's help.

Dahmas (1970) concluded that "It is common to classify problems as the three percent problems, mixture problems, work problems, and others. But, classifying problems encourages students to look for relatively unimportant common characteristics. At best, most of the classifications serve only as temporary crutches for a few students."

Dahmas (1970) devised a method for solving word problems. He recommended a direct translation of the problem statement into a mathematical statement. He found that this method resulted in better performance than did Polya's method of heuristically solving the
whole problem. However, as Bassler (1975) points out, the problems used by Dahmas were very easily translated into mathematical sentences and therefore, were not representative of many word problems. When Bassler compared Polya's method to the Dahmas method on a different set of problems, the results indicated that a correct equation was generated more often using Polya's method.

READING ABILITY

The reasons that a familiar appearance in a story problem reduces the difficulty level is unclear; but reading ability seems to be a factor. Call (1966) examined some special reading problems encountered by students reading mathematics. He recommended special reading instruction as a means of improving student performance on word problems. Jones (1971) has designed several reading text books used in elementary schools. She has included sections in her reading texts designed to teach the special reading skills required in the "content" areas. She has made specific recommendations for reading mathematical word problems. One suggestion was to look for small changes in the wording of problems which could change the problem's meaning. Another suggestion was to focus on the relevant information necessary to solve the given problem.
Some research has indicated that subjects may form a "story schema" for word problems. Mandler and Johnson (1977) studied the relation of story structure and recall. Their results implied that subjects form a "story schema" and "use this schemata to guide comprehension during encoding and as a retrieval mechanism during recall." Anderson (1977) and Minsky (1975) also found subjects' schemata provided a framework which might aid students in comprehension of specific readings.

Aiken (1973) reported a high correlation between reading ability and mathematical problem solving; but few studies have investigated the relationship between prior familiarity with the problem form, formation of story schema, and reading ability. Filano (1957) determined that a student's ability to read and understand mathematical explanations was quite good after he had received instruction on highly similar material.

RELATED RESEARCH

Many studies may not have recognized the effect of previously encountered problems which use the same problem appearance as the criterion test items. Mager (1962) in his work on behavioral objectives, identified three components necessary to specify an "instructional" objective. The components were the behavioral performance, the conditions under which this performance will be
observed, and the criterion for acceptable performance. Mager stressed the need to state these objectives clearly. However, educators have frequently described the conditions for the instructional objective in such detail, that the student was given the test item—except for the numbers. Often objectives were written as if each problem required a unique instructional objective and has nothing in common with any other instructional objective. Schoen (1976) criticized many individualized programs because they were based on instructional objectives which were so narrow that only low level skills were required. On the surface, these criticisms appear to be supportive of the theory that providing students with "sample" test items reduces the cognitive level required to solve those problems. The results of several studies (Kulic, 1974; Wagner 1973; Kipps, 1970; Wood, 1972) could be attributed to the use of "sample" test items. Wood (1972) reported that providing students with "sample" test items resulted in improved performance on tests. Kipps (1970) allowed students to retake "equivalent" tests until they "mastered" the material. If previous exposure to "sample" test items does in fact reduce the cognitive level required to solve the given problem, then problems which employ an unencountered appearance may be necessary to measure transfer.
SUMMARY

The review of the literature has indicated that students have difficulty in transferring knowledge to solve new problems. This study may show that many students have difficulty transferring knowledge to solve problems which are structurally identical to previously encountered problems. If varying only a few irrelevant properties markedly affects student performance, this may explain in part, why students are unable to transfer knowledge to solve new problems.

Perhaps many students are able to solve problems only after they have previously solved equivalent problems, which were presented using the same set of contextual details and other irrelevant properties. Parents and teachers have often asked why students do well in school and then cannot apply their knowledge to practical problems? Perhaps some students have only been evaluated on problems which were presented using a familiar form and problem appearance. This experience enabled the students to operate successfully at the memorizational level.

Merrill's model has defined the cognitive levels required to correctly classify or solve problems in terms of previously unencountered instances. Few other cognitive models have considered the student's problem solving experience carefully; yet experience may be among the most influential factors.
Many educators would agree that one long range objective is to develop the ability to transfer knowledge to previously unencountered problems. If increasing the variability of the irrelevant attributes in homework and example problems can be shown to promote transfer, there are some immediate implications for constructing problem sets. Research must investigate the long range effects of variation in the problem forms encountered during instruction. Textbooks must provide more variability among the irrelevant attributes associated with the problem sets. If the problem form is found to be a significant factor, then students probably use the irrelevant properties to help reduce the difficulty levels for the problems which are matched to previous homework problems.

The literature has expressed a need for empirical studies to determine the effect of problem appearance on problem difficulty. Merrill's model for cognitive levels suggests that the form of problems encountered by students is a powerful factor. Tennyson, Markle and others have made specific suggestions concerning how sets of instructional items should be constructed. But studies which directly confirm their conjectures are almost non-existent. Even if these theories prove to be accurate predictors, the causes for the observed differences are not obvious. Many factors such as reading
ability, story schemata, order of data, and memory may interact to increase or decrease the difficulty of specific problems. Thus the existing research provides some conjectures about familiarity with problem appearance and problem difficulty; but there is a lack of empirical verification of these conjectures.
CHAPTER III

METHODS AND PROCEDURES

The purpose was to determine the effect of prior familiarity with contextual details and irrelevant properties affecting the problem appearance on the difficulty of selected word problems in arithmetic. Some precise definitions may be helpful.

<table>
<thead>
<tr>
<th>Problem Type:</th>
<th>Two word problems are of the same problem type if the problems can be solved using the same computational algorithm.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem Form:</td>
<td>Two word problems are of the same problem form if the irrelevant properties which affect the problem appearance are as similar as possible.</td>
</tr>
<tr>
<td>Relevant Property:</td>
<td>Two problems differ by relevant properties if they are of different problem types.</td>
</tr>
<tr>
<td>Irrelevant Property:</td>
<td>Two problems differ by irrelevant properties if they are of the same problem types.</td>
</tr>
</tbody>
</table>
Prior Familiarity: The student is familiar with a problem if he has previously solved a problem which used the same problem form.

Convergent: Two instances of the same problem type are convergent when the irrelevant attributes are as similar as possible.

Divergent: Two instances of the same problem type are divergent when the irrelevant attributes are as dissimilar as possible.

Matched: A matched relationship exists between instances of two different problem types when the irrelevant attributes are as similar as possible.

Not Matched: Instances of two different problem types are not matched when the irrelevant attributes are as dissimilar as possible.

Errors of Misclassification: An error of misclassification is an instance when a subject applies an algorithm which is incorrect but which would have correctly solved
a problem of a different problem type.

OVERVIEW

This study consists of two separate experiments. Experiment I was designed to document that the problem appearance is frequently associated with the method or algorithm used to solve a given problem. In particular, when the problem appearance is not systematically varied, misconceptions are likely to result. Experiment II was designed to answer two questions. First, is a problem which has a familiar appearance less difficult for a student than a structurally equivalent problem with an appearance that has not been encountered by the student during the instructional sessions? Second, will systematic variation of problem form result in improved performance on equivalent problems with an unencountered problem appearance?

The subjects for Experiment I were fifth grade students from two Columbus area elementary schools. Experiment I was conducted during May, 1978. The subjects for Experiment II were sixth grade students from three Columbus area schools. Experiment II was conducted during the autumn of 1978.
Since the two experiments were related, no schools were chosen to participate in both experiments. Therefore, interactions which might have resulted from subjects participating in both experiments have been avoided.

HUMAN SUBJECTS APPROVAL

A request for the use of human subjects in both experiments was presented to the Chairperson of the Human Subjects Review Committee for The Ohio State University. Approval to use human subjects was granted on May 5, 1978.

EXPERIMENT I

SETTING

Experiment I was conducted at two Columbus area elementaries during May, 1978. The subjects were 85 fifth-grade students. The students were instructed in small groups of seven subjects. Fifth-grade students were chosen for several reasons. First, the subjects generally had not encountered word problems similar to those used during the study. Secondly, the subjects were from a district that was at or above the national average in achievement so that many students would be able to solve the word problems after a modest amount of instruction. Thirdly, due to the flexibility of the
participating classes, it was possible to work with small
groups of students without seriously disrupting the
existing educational programs.

SELECTION OF SUBJECTS

Two elementary schools were selected to participate
in Experiment I. There were two fifth grade classes in
each of the participating schools. Therefore, four classes
of fifth grade students were selected as subjects. In
each class, subjects were randomly assigned to create
three smaller groups of seven students. The small group
was the experimental unit of analysis. Two of the three
small groups in each class were assigned to an instruction­
ional treatment. The two instructional groups were given
instruction on two days for thirty minutes each day.
The third group was a control and did not receive any
instruction. On the third day the criterion test was
administered to the entire class. The independent
variable in this experiment was the set of problems solved
during instruction.

ITEM CONSTRUCTION

Two problem types and two problem forms were con­
structed for Experiment I. Each of the problems presented
to the subjects contained four bits of numerical data.
The numerical information given in the problem statement
and an algorithm which correctly solves each problem type are described as follows.

Type I  Given: a, b, x, and y.

Algorithm: \( S = ax + by \)

Type II  Given: a, b, x, and y.

Algorithm: \( S = ax + (b - a)y \)

Examples of each of the Form A combinations are listed as follows. Important changes which distinguish the different problem types have been underlined.

Type I - Form A

Two men left Columbus in a car on a long trip. One man drove at 40 miles per hour for 7 hours. The other man drove at 50 miles per hour for 12 hours to complete the trip. How far was the trip?
Type II - Form A

Two men left Columbus in a car on a long trip. One man drove at 40 miles per hour for 7 hours. The other man drove at 50 miles per hour so that the trip was completed in 12 hours. How far was the trip?

Examples of each of the Form B combinations are listed below. Important changes which distinguish the different problem types have been underlined.

Type I - Form B

A man accepted a job where he was paid $7 each day that he was training. His pay was then increased to $15 for each day after his training period had finished. His training period lasted for 6 days. After his training, he worked 18 days. How much did he earn on this job?

Type II - Form B

A man accepted a job where he was paid $7 each day that he was training. His pay was then increased to $15 for each day after his training period had finished. His training period lasted for 6 days. Including his training, he worked 18 days. How much did he earn on this job?
There was an effort to match the irrelevant properties within a given problem form. Many of the similarities among the contextual details for two instances of the same problem form are obvious. For example, both Form A problems involve a car, two drivers, time, and distance traveled. Each problem began, "Two men left Columbus in a car on a long trip." The question was, "How far was the trip?" The relevant change that distinguishes a Type I - Form A problem from a Type II - Form A problem has been underlined in each item. A Type II problem can be derived from a Type I problem by replacing the phrase, "for 12 hours to complete the trip?" with the phrase, "so that the trip was completed in 12 hours." Although the problem appearances for the two problems are highly similar, the algorithms needed to correctly solve the problems are different.

The two Form B problems may at first glance appear to be identical. Both problems began, "A man accepted a job where he was paid $7 for each day that he was training." The rate of pay was then changed after the training period ended. Four numbers were presented in each problem statement. The only difference between the Type II and Type I problem statements was that the word "including" was used to replace the word "after" in the
Two forms of the same problem type were made to appear dissimilar. To accomplish this objective, many irrelevant properties were changed so two problems of the same problem type appeared to be dissimilar. First, different contextual details were used to derive each form. Form A problems involved a car, distance traveled, and time. Form B problems involved pay for a job, the number of days worked, and a training period. The order in which the data was presented differed for the two problem forms. Type I problems were solved using the algorithm $S = ax + by$. The order of the data in Form A problems was $x$, $a$, $y$, and $b$. The order for the data in Form B problem statements was $x$, $y$, $a$, and $b$. Therefore, to solve a Type I - Form A problem, one must multiply the first and second numbers and then add the product of the third and fourth numbers. To solve a Type I - Form B problem, one multiplies the first and third numbers and adds this result to the product of the second and fourth numbers.

One objective was to determine if students formed misconceptions, as predicted by Merrill (1973) and Tennyson (1974a), when they were presented with sets of examples which were convergent and not matched. Therefore, the irrelevant properties were not matched when
deriving the different problem types used during the instructional treatments. However, the instances of each problem type were matched as closely as possible so that the instances of each problem type were convergent.

A set of nine problems using Type I - Form A, and nine problems using Type II - Form B were presented during Treatment AB. Treatment BA subjects encountered a set of nine Type I - Form B and nine Type II - Form B problems. Both treatments consisted of examples which were convergent and not matched. Therefore, it was predicted that some students would associate the algorithm used to solve the problems with the contextual details and problem forms that they had encountered during their instructional sessions.

THE INSTRUCTIONAL SESSIONS

The experimental groups were instructed in a room separated from the remaining members of their class. The students were seated at two small tables facing a chalk board. The small groups each received instruction in two sessions which lasted approximately thirty-five minutes each. The researcher wrote one example of each of the two problem type-form combinations on the board prior
to each session.

When the subjects entered the room for the first session they were given two problem sheets. Each problem sheet contained three problems. The first problem was then read to the small group by the researcher. The researcher then discussed the problem with the members of the small group until a valid solution was determined. A sample of the interaction between the researcher and the small group members is presented in Appendix E. After the first problem was completed, the subjects were asked to solve the remaining two problems on the first sheet.

The first problem on the second sheet was then read to the subjects. Another discussion between the researcher and the subjects took place which led to a solution. The subjects were then directed to solve the remaining two problems on the second sheet. The researcher acted as a tutor while the students completed the problems.

The second day the subjects were given a total of 12 problems on four separate sheets. These 12 problems consisted of six problems using each of the two problem type-form combinations derived for their treatment. The researcher guided a solution to the first problem of each type. The subjects were then allowed to solve the
remaining 10 problems at their own rate. The researcher checked the results for each student, informing the subjects which problems had been solved correctly, and correcting any errors.

THE CRITERION TEST

The criterion test consisted of 12 items. There were three problems for each of the four problem type-form combinations. That is, three problems using Type I - Form A, three problems using Type I - Form B, three problems using Type II - Form A, and three problems using Type II - Form B. A copy of the criterion test can be found in Appendix B.

An effort was made to reduce any effect which might be due to the order of the items found on the test. To accomplish this, the pages were not in the same order for each test. Each page appeared as the first page on one-fourth of the tests, as the second page on one-fourth of the tests, as the third page on one-fourth of the tests, and as the last page on one-fourth of the tests.

The tests were administered to intact classes. The following directions were verbally given to the subjects.
1. Show all your work in the space provided below each problem statement.

2. Read each problem carefully.

3. If you have difficulty with one problem you may choose to attempt a different problem.

Many students completed the twelve item test in roughly 15 minutes. However, the subjects were allowed about 35 minutes and in a few cases a little longer so that they could complete the items. Only a few subjects needed more than 30 minutes to complete the 12 item test.

VARIABLES

The criterion test consisted of 12 items, three instances for each of the four problem type-form combinations. The problems on the criterion test fell into two categories. These two categories constituted the two variables. One variable was estimated by the performance for problems which matched the problem type-form combinations presented in Treatment AB. The other variable was estimated by the performance on problems which employed the same problem type-form combinations presented during
Treatment BA. Therefore, the criterion test was considered as two six-item subtests for the purpose of analysis. Test AB consisted of three Type I - Form A and three Type II - Form B problems. These problems matched the type-form combinations encountered by Treatment AB subjects. Test BA consisted of three Type I - Form B and three Type II - Form A problems. Thus, the problems appearing on Test BA matched those presented to the subjects during Treatment BA. Hence, Test AB and Test BA estimated the two variables for Experiment I.

GRADING OF THE TEST

Each problem was graded as either correct or incorrect. The criterion for scoring a problem as correct was that a valid algorithm was applied. Problems which contained only minor computational errors but which used a correct algorithm were scored as correct. To estimate the "errors of misclassification," it was also noted if a subject applied a Type I algorithm to solve a Type II problem, or if a subject applied a Type II algorithm to solve a Type I problem.

DEBRIEFING

The researcher reported the results of the study to each class a few days after the class completed the testing. During this discussion, the hypotheses of the
research and the differences between the treatments were explained. An effort was made to clarify any misconceptions which the subjects may have formed, and to illustrate how irrelevant attributes may interfere with their performance.

LIMITATIONS

The small groups were each assigned seven students prior to the treatments. However, there were absentees which prevented some students from participating in the treatments or completing the testing. Three subjects missed the testing and were excluded from the analysis. One subject missed both experimental sessions and was excluded from the experiment. Two subjects missed one instructional session and were excluded from the analysis. Nine others were excluded at random to yield four small groups with six subjects for each of the three treatments.

One class had a teacher who had frequently worked word problems with her class. She had stressed having her students translate the problems and solve. She frequently recommended that the subjects look for small differences in problem statements as recommended by Jones (1971). The mean performance of this class (noted as class #1 in Table 6) was generally above the corresponding averages for the other three classes.
However, the interaction between the treatments on the two sections of the criterion test remained consistent with the results for the other classes.

The small groups were formed by randomly assigning subjects within a class to treatments. The method used to assign subjects to small groups did not account for the differences between the ability levels of the different classes. But this weakness could not be avoided since the alternate methods of assignment of subjects to groups which would have mixed subjects from different classes within the same group was operationally not feasible.

EXPERIMENT II

SETTING

The experiment was conducted in three Columbus area schools during the Autumn of 1978. The subjects consisted of nine sixth-grade classes, three from each of the participating schools. Sixth grade students were chosen as a source of subjects for several reasons. First, the subjects would generally not be able to solve the word problems used in this experiment prior to instruction. Second, sixth grade students have rarely encountered problems which require several steps to solve. Thus the problems used in the treatments were essentially unencountered. Third, sixth grade students at the beginning of the school year were similar to the
fifth grade students used near the end of the school year for Experiment I. Fourth, since the students from these schools have scored at or above the national norms on achievement and IQ tests, many subjects would learn to solve these problems correctly after a modest amount of instruction. Finally, the students were accessible to the experimenter. Due to the flexibility of the existing programs in the participating schools it was possible to instruct the small groups without seriously disrupting the existing curriculum.

SELECTION OF SUBJECTS

Three Columbus area schools agreed to participate in Experiment II. There were three sixth-grade classes in each of the schools thus, a total of nine classes were selected.

The participating classes were divided into either three or four small groups of seven students depending on the class size. To assure that the small groups were relatively equal in ability, results from the Mathematics Applications Section of the Stanford Achievement Test (1971) were used to stratify the different classes. Each class was first divided into high and low ability subjects as determined by the aptitude scores. Three high ability and three low ability students were assigned at random to each of the small groups. In addition, one
subject was assigned at random to each small group to serve as an alternate. The results for this subject were excluded from the analysis of the small group unless it was necessary to use the alternate to replace a subject who may have missed either the instructional sessions, or the testing.

A total of 29 small groups were used in this study. At the first school there were nine small groups. Two small groups were assigned to each of the three instructional treatments and three groups were assigned to the control treatment. The three sixth-grade classes at the second school were divided into eleven small groups. Three small groups received each instructional treatment and two small groups were control groups. The three sixth-grade classes at the third school were assigned to nine small groups. Two groups received each of the three instructional treatments and three groups were control groups. Therefore, of the 29 small groups, seven groups were assigned to each instructional treatment, and eight groups were assigned to the control treatment.

ITEM CONSTRUCTION

Experiment II was designed to answer two questions. First, is a problem which uses a familiar appearance less difficult than a mathematically equivalent problem which uses contextual details that have not been
encountered by the subject? Second, when the problem forms are systematically varied during the instructional sessions, does this result in better performance on equivalent problems with an unencountered appearance? Three problem types and four problem forms were derived. Thus a total of 12 problem type-form combinations were constructed for Experiment II. The problem forms were denoted as Form A, Form B, Form C, and Form D. Three instructional treatments were derived for Experiment II. The three treatments were denoted as Treatment AAA, Treatment BBB, and Treatment ABC. The difference between the instructional treatments was determined by the problem forms used to present the set of problems during each session.

Problems which used the same problem form were constructed to be as similar as possible. Because the problems within a problem form are of different problem types, there was a limit to how similar the problem appearances could be matched. But many similarities within each problem form are obvious. In order to make the problems of the same problem form appear similar to each other, each problem began with the same phrase. The wording of each problem within a problem form was made to be as similar as possible. Examples of each problem type-form combination are listed below.
Type I - Form A

Two men left Columbus in a car on a 360 mile trip. One man averaged 44 miles per hour for 6 hours. What was the average speed of the other driver if the trip was completed in 9 hours?

Type II - Form A

Two men left Columbus in a car on a long trip. One man drove 32 miles per hour for 4 hours. The other man averaged 52 miles an hour for 6 hours. How far was the trip?

Type III - Form A

Two men left Columbus in a car on a long trip. One man drove 32 miles per hour for 4 hours. The other man drove 52 miles an hour for 6 hours. What was the average speed for the trip?

Type I - Form B

There are 16 boxes on a loading dock, some are red and the rest are green. The total weight of the boxes is 270 pounds. If each of the 6 red boxes weighs 15 pounds, how much does a green box weigh?
Type II - Form B

There are 6 red and 9 green boxes on a loading dock. Each of the red boxes weighs 35 pounds. Each of the green boxes weighs 20 pounds. What is the total weight for the boxes?

Type III - Form B

There are 6 red and 9 green boxes on a loading dock. Each red box weighs 35 pounds. Each green box weighs 20 pounds. What is the average weight for a box on the loading dock?

Type I - Form C

A bowler had a 212 average per game for the 10 games he bowled on Saturday in a weekend tournament. He finished his bowling on Sunday with a total of 5680 for the 30 game tournament. What was his average on Sunday?

Type II - Form C

A bowler was in a weekend tournament. He averaged 210 for each game he bowled on Saturday and 225 for each game on Sunday. He bowled 16 games on Sunday and only 8 games on Saturday. What was his total score for the tournament?
Type III - Form C

A bowler was in a weekend tournament. He averaged 210 for each game he bowled on Saturday and 225 for each game on Sunday. He bowled 16 games on Sunday and only 8 games on Saturday. What was his average score for the tournament?

Type I - Form D

A man accepts a job scraping and painting a house. He is paid $22 a day for the 10 days that he scrapes the house. To both scrape and paint the house takes 16 days. He is paid $400 for this job. How much is he paid for each day that he paints the house?

Type II - Form D

A man accepts a job scraping and painting a house. He is paid $15 each day for scraping the house and $35 each day that he is painting. He works 6 days scraping and paints for 6 days. How much will he have earned on this job?

Type III - Form D

A man accepts a job scraping and painting a house. He scrapes the house for 6 days and is paid $15 a day. His pay is then changed to $25 per day for the 4 days that he paints the house. What is the average daily pay for
Two problems which employed the same problem form, but different problem types, were matched as closely as possible on the irrelevant properties. These irrelevant properties included the contextual details, the order of the data, and the use of certain key words. For example, consider the following:

Type II - Form A

Two men left Columbus in a car on a long trip. One man drove at 32 miles per hour for 4 hours. The other man averaged 52 miles an hour for 6 hours to complete the trip. How far was the trip?

Type III - Form A

Two men left Columbus in a car on a long trip. One man drove 32 miles per hour for 4 hours. The other man drove 52 miles an hour for 6 hours. What was the average speed for the trip?

Both of these Form A problems involved a car, two drivers, distance traveled, and time. Each problem began, "Two men left Columbus in a car on a long trip." The order of the data presented in the problem statements was the same. Also, key words such as "average" were used in each problem statement. The change that distinguishes the Type II - Form A problem from the Type III -
Form A problem was the question. In the Type II problem the question was, "How far was the trip?" For the Type III problem the question was, "What was the average speed for the trip?"

Two problems which were of the same problem type but different problem forms, were not matched on several irrelevant properties. Consider the Type I - Form A and Type I - Form B problems given earlier.

Type I - Form A

Two men left Columbus in a car on a 360 mile trip. One man averaged 44 miles per hour for 6 hours. What was the average speed of the other driver if the trip was completed in 9 hours.

Type I - Form B

There are 16 boxes on a loading dock, some are red and the rest are green. The total weight of the boxes is 270 pounds. If each of the 6 red boxes weighs 15 pounds, how much does a green box weigh?

Both problems can be solved using the same algorithm $S = (T - ax)/(c-a)$. The contextual details of the Form A problem involved a car, two drivers, two speeds, and the time for each driver. The details of Form B problems include weight, the number of boxes on a loading dock, and the colors of the boxes. The order in which the data
was presented differed for the two forms. The order of
the data for Form A problems was T, x, a, and C. The
order in which the data was presented in Form B problems
was c, T, a, and x.

The numerical data which was given in each problem
statement and an algorithm which solves each problem
type are described below.

Type I:  Data Given - a, c, x, and T
        Algorithm - \( S = \frac{T - ax}{c - a} \)

Type II: Data Given - a, x, b, and y.
         Algorithm - \( S = ax + by \)

Type III: Data Given - a, x, b, and y.
          Algorithm - \( S = \frac{ax + by}{a + b} \)

The contextual details in each problem statement and
the order in which the data was presented were varied to
derive four different forms for each of the three problem
types. The topic used for each problem form were as
follows:

Form A: The problem statements involved the speed and
distance traveled by two men in a car.
Form B: The problems involved the weights of several boxes on a loading dock.

Form C: The problems involved the scores for a bowler who participates in a two-day tournament.

Form D: The problems involved a man who was paid for scraping and painting a house.

Figure 3 indicates the order in which the data was presented for each type-form combination.

<table>
<thead>
<tr>
<th></th>
<th>Type I</th>
<th>Type II</th>
<th>Type III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form A</td>
<td>T, x, a, c</td>
<td>x, a, y, b</td>
<td>x, a, y, b</td>
</tr>
<tr>
<td>Form B</td>
<td>c, T, a, x</td>
<td>a, b, x, y</td>
<td>a, b, x, y</td>
</tr>
<tr>
<td>Form C</td>
<td>x, a, T, c</td>
<td>x, y, b, a</td>
<td>x, y, b, a</td>
</tr>
<tr>
<td>Form D</td>
<td>x, a, c, T</td>
<td>x, y, a, b</td>
<td>a, x, y, b</td>
</tr>
</tbody>
</table>

FIGURE 3

ORDER OF THE DATA FOR EACH TYPE-FORM COMBINATION
There were three instructional treatments. Each treatment was presented to small groups of seven subjects. There were three sessions of roughly twenty-five minutes for each instructional treatment. During each session two instances of each of the three different problem types were solved. The problem form used to derive the problems presented during the instruction was the independent variable.

Treatment AAA presented Form A problems during each of the three sessions. Treatment BBB presented Form B problems during each of the three sessions. During Treatment ABC, Form A problems were presented the first day, Form B problems on the second day, and Form C problems were presented on the third day. There were three problems on each of the six pages used during the three instructional sessions. To prevent the students from associating the correct algorithm used to solve a problem with the position of a problem on the page, the order of the problem types on the pages were varied. Since the number of permutations of three items is equal to six and there are six pages of three problems, each of the six possible orders was used on one problem sheet for each treatment. The items used in each treatment are presented in Appendix C.
THE INSTRUCTIONAL SESSIONS

Each small group was instructed in a classroom or area separate from the remaining members of their class. The students were seated at desks or tables facing a small chalk board. During each session six problems were discussed and solved. The subjects were given a sheet with three problems at the start of each session. The subjects were asked to read the first question and attempt to solve the problem. After one minute, one of the students was asked to read the question aloud. An interaction between the researcher and the small group ensued, which led to the solution. For a more detailed description of a typical interaction between the researcher and the subjects see Appendix E. The second and third problems were then solved using a similar instructional sequence.

The second page of problems was given to the subjects after the first page had been completed. The researcher then requested that each student attempt to solve the problems on the second page. When a student had finished a problem the researcher checked the answer, if the answer was correct the student proceeded to the next problem. If the student was incorrect or had trouble with the solution, the researcher acted as a tutor helping the subject solve the problem. Because there were only seven students in a small group the researcher was able to
handle the questions. Occasionally, when several students experienced a common difficulty, the researcher discussed the problem with the whole small group in the manner described in Appendix E.

The sessions were completed when each subject finished the six problems. The duration of a single session was generally between 20 and 30 minutes.

CRITERION TEST

The criterion test consisted of 24 items. Two items that used each of the twelve problem type-form combinations were derived for the test. There were six items which used each of the four problem forms. The test was administered as two twelve-item tests which each lasted thirty-five minutes. Copies of the tests are attached in Appendix D. Each twelve-item test was four pages long. Each page consisted of three problems, one of each of the three problem types. The order of the problems on each page of the test was determined by random assignment.

The pages for the different copies of the test were ordered so that each page appeared as the first page on one-fourth of the tests, as the second page on one-fourth of the tests, as the third page on one-fourth of the tests, and as the final page on one-fourth of the tests.
The first twelve-item test consisted of items which used Form A and Form B. The second twelve-item test consisted of items which used Form C and Form D.

Each of the two twelve-item tests was administered to the class as a whole rather than to the individual small groups within the classes. The first test was given the day following the third instructional session. The second test was administered as soon after the first test as possible. This was usually the next day, but on three occasions it was two days later. The subjects were given the following directions prior to attempting the test.

1. Show all of your work in the space provided under each problem.

2. Read each problem carefully.

3. You may find some problems more difficult than others. If you are not able to solve a problem you may move on to another problem.

Many subjects completed the twelve-item tests in roughly fifteen minutes. However, the subjects were allowed about thirty-five minutes so that all the subjects could complete the problems which they were able to solve. Only a few students used more than thirty minutes during a testing session.
SCORING OF THE TEST

Each item on the test was graded as correct or incorrect. The item was scored as correct if the algorithm used was valid. Therefore, problems which had only minor computational errors were scored as correct. Problems which applied an algorithm which was not valid were scored as incorrect.

DEBRIEFING

The researcher reported the results of the study to each class a few days after the class completed the testing. During this discussion, the hypotheses of the research and the differences between the treatments were explained. An effort was made to clarify any misconceptions which the subjects may have formed, and to illustrate that irrelevant attributes may interfere with their performance.

VARIABLES

There were four variables estimated in Experiment II. Each of the variables was estimated by the performance on problems using a specific problem form. There were three problem types and four problem forms. The criterion test consisted of twenty-four items. There were two items which employed each of the twelve problem type-form combinations. The four subtests were not
based on the problem type but on the problem form. There were six items which used each of the four problem forms.

The four variables were estimated by performance on the four particular problem forms. The four problem forms were denoted as Form A, Form B, Form C, and Form D. These were estimated by scores for Test A, Test B, Test C, and Test D. Each subtest consisted of six problems which used a specific problem form. For instance, Test A consisted of six problems which employed Form A.

LIMITATIONS

Each group was assigned six subjects plus one alternate subject who was automatically excluded from the analysis unless there were absentees. Four subjects who missed one of the three instructional sessions were instructed individually by the researcher. These four subjects worked the same problems which they had missed due to their absence. Three students missed the criterion test. One of these students received a make-up test the following day. The other two were not available for a make-up test and therefore, were excluded from the analysis. They were replaced in the analysis by the two subjects who served as alternates during the treatments.
The small groups were formed by randomly assigning subjects within a class to the treatments. This method of assignment may result in the small groups within the treatments having slightly different abilities. But because the class medians on the Stanford Achievement Test (1971) ranged from 27 to 31 this problem was not viewed as serious. Alternate methods of assignment were rejected because of the organizational problems which prevented instructing small groups containing students from several different classes.

The test was administered on two different days. All of the Form C and Form D problems appeared on the second day. Therefore, there was the possibility that the subjects had either learned or forgotten solution techniques between the two testings. This effect was considered less serious than exposing the subjects to all four forms of the problem types on the first day and then testing those students on problems using the same form again on the second day. The test was considered to be too long to administer on one day.

A few subjects became frustrated by the difficulty of the problems and essentially gave-up on the test. This may have been reflected in their results, but most students worked very hard during the instructional treatments and testing sessions. In addition, most students were interested in their performances.
SUMMARY

Many students display very little ability to transfer knowledge to solve problems. Prior familiarity with specific problem appearance may be a factor that reduces the difficulty level of problems that employ a similar problem appearance. If this is true, one would expect that many students use a familiar problem appearance as a "crutch" to help themselves determine how to solve the given problem. The first experiment investigated how problem appearance might be associated with the solution algorithm. The problem sets presented during the two instructional treatments were convergent and not matched. Therefore, it was likely that many subjects might be tricked into forming misconceptions by associating the problem form with an algorithm. Two variables were measured in Experiment I. One variable represented performance on problems which matched the problem type-form combinations encountered during instruction. The second variable estimated performance on problems whose type-form combinations were not matched to the problems encountered during the instructional sessions.

In the second experiment it was predicted that subjects would perform better on those problems which used a familiar problem appearance. It was also predicted that when subjects encountered a greater variety of problem appearances during the instruction that they
would perform better on an unencountered problem form. Four problem forms were derived for each of the three problem types. Any problems of the same problem type could be solved using the same algorithm. Three instructional treatments were designed for Experiment II. Two of the treatments did not vary the problem form, the third instructional treatment varied the problem form. Four variables were measured. Each of the variables represented performance on one of the four problem forms. Different subjects had encountered different problem forms during the instruction. Therefore, it was predicted that the subjects would score higher on those problems which used the problem forms that had been encountered during the instruction. A multivariate analysis followed by appropriate univariate procedures were applied to the scores for the different problem forms.
CHAPTER IV

DATA ANALYSIS

DESIGN AND HYPOTHESES

The subjects participating in Experiment I received one of three treatments. Two criterion variables were measured for each subject. The three levels for the independent variable were denoted as Treatment AB, Treatment BA, and the Control Treatment. The small groups receiving Treatment AB solved Type I - Form A and Type II - Form B problems. The groups receiving Treatment BA solved Type I - Form B and Type II - Form A combinations. The small groups which did not receive any instruction were denoted as the Control Treatment. Estimates for the two variables were obtained using the results from Test AB and Test BA. Test AB consisted of three Type I - Form A and three Type II - Form B problems. Test BA consisted of three Type I - Form B and three Type II - Form A problems. Figure 4 presents the design used for Experiment I.
Treatment  | Test AB | Test BA
-------|--------|--------
AB      | X      | Y      
BA      | X      | Y      
CONTROL | X      | Y      

**FIGURE 4**

**DESIGN FOR EXPERIMENT I**

The hypotheses were to be tested using a multivariate analysis of variance followed by appropriate univariate procedures determined by the results for the multivariate analysis. The hypotheses were as follows:

1. There were differences between the scores on Test AB for the three treatments.
2. There were differences between the scores on Test BA for the three treatments.
3. There were differences between the number of "errors of misclassification" on Test AB for the three treatments.
4. There were differences between the number of "errors of misclassification" on Test BA for the three treatments.

There were 81 subjects who completed the experiment. Prior to the analysis nine subjects were randomly excluded. These exclusions created four small groups of
six subjects for each of the three treatments. The statistical tests that were employed made the assumption of equal cell variances and yet remained robust even when there were unequal variances if the cell sizes were equal. The loss of data which resulted from the random exclusion of the subjects was considered minimal. This was particularly true since the unit of analysis was the small group and no small groups were excluded.

ANALYSIS

The criterion test consisted of two subtests denoted as Test AB and Test BA. Test AB was composed of three Type I - Form A problems and three Type II - Form B problems. Subtest BA consisted of three Type I - Form B problems and three Type II - Form A problems. The items were scored as either correct or incorrect. A problem was scored as correct if a correct algorithm was applied to solve the problem. It was believed that many students might apply the Type II algorithm to solve some Type I problems, and conversely. To obtain an estimate of the "errors of misclassification" each instance where a subject applied a Type II algorithm to solve a Type I problem or applied a Type I algorithm to solve a Type II problem was noted. Therefore, each test was scored in two ways. One score indicated the number of problems that each student solved correctly. The other score indicated the number of "errors of misclassification."
A multivariate analysis of variance was applied to determine if there were differences between the three treatments on Test AB and Test BA.

A multivariate analysis of variance was conducted on the Ohio State University IBM 370 Computer using the Clyde (1969) Manova Program.

The KR-20 estimates of reliability were found to be .84 for Test AB, and .94 for Test BA. Both of these tests were six-item tests. The reliability estimate for the twelve-item test formed by combining Test AB and Test BA was found to be .79. This reliability was apparently lower than the subtest reliabilities even though most sources such as Glass and Stanley (1970) state that increasing the number of items will result in an increase in the reliability estimates. However, one assumption they made was that all the items in the test were estimating the same variable. But in this case the two tests were designed specifically to measure two different variables. Thus a decrease in the reliability resulting from the combining the two tests might be expected.

The small group was the experimental unit of analysis. The means and standard deviations for the small group averages on Test AB and Test BA are presented in Table 1.

The results from the multivariate analysis of variance are presented in Table 2. The existence of
TABLE 1
MEANS AND STANDARD DEVIATIONS FOR SMALL GROUP
AVERAGES ON TEST AB AND TEST BA

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Test AB</th>
<th>Test BA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>AB</td>
<td>5.248</td>
<td>.290</td>
</tr>
<tr>
<td>BA</td>
<td>2.207</td>
<td>.895</td>
</tr>
<tr>
<td>CONTROL</td>
<td>2.747</td>
<td>.833</td>
</tr>
</tbody>
</table>

large F-ratios for the multivariate test justified
subsequent applications of univariate tests for each of
the variables individually. The results for the uni-
variate tests are also reported in Table 2.

TABLE 2
MULTIVARIATE AND UNIVARIATE ANALYSIS OF VARIANCE
SCORES FOR VARIABLES TEST AB, TEST BA,
USING WILKS LAMBDA CRITERION

<table>
<thead>
<tr>
<th>Variables</th>
<th>Test</th>
<th>df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test AB, Test BA</td>
<td>M</td>
<td>(4,16)</td>
<td>35.040</td>
<td>.001</td>
</tr>
<tr>
<td>Test AB</td>
<td>U</td>
<td>(2,9)</td>
<td>19.988</td>
<td>.001</td>
</tr>
<tr>
<td>Test BA</td>
<td>U</td>
<td>(2,9)</td>
<td>123.517</td>
<td>.001</td>
</tr>
</tbody>
</table>

The multivariate analysis indicated differences
between the three treatments. In addition, the subse-
quently univariate tests also indicated that differences
existed (p < .001) between the treatments on both Test
AB and Test BA. To better understand the nature of
these differences, plots of the cell means are presented
in Figure 5.

FIGURE 5
PLOT OF TREATMENT MEAN SUBTEST SCORES

Any pairwise differences between the treatment averages were of interest. Therefore, Tukey's test for honest significant differences between treatment means was applied. For Test AB a difference of 1.43 ($p < .05$) or 1.864 ($p < .01$) was required to conclude that two treatment means were different. Thus, there were two significant comparisons ($p < .01$). First, the difference of 3.041 favoring Treatment AB over Treatment BA. Second, the difference of 2.501 favoring Treatment AB over the Control Treatment. But the difference of .540 between the Control Treatment and Treatment BA was not significant ($p < .05$).
On Test BA, Tukey's test required a difference of .741 \( (p < .05) \) or .965 \( (p < .01) \) between two means to conclude that the treatments differed on Test BA. Again, two pairwise comparisons were significant \( (p < .01) \). Treatment BA scored higher than Treatment AB because the difference between the cell means was 3.792. The difference of 3.377 favoring Treatment BA over the Control Treatment was also significant. But the difference of .415 between the Control Treatment and Treatment AB was less than .741 and therefore, not sufficient to conclude that Treatment AB and the Control Treatment differed on Test BA \( (p < .05) \).

The averages on Test AB and Test BA for the twelve small groups which participated in Experiment I are presented in Table 3.

Each test was also graded for "errors of misclassification." An "error of misclassification" was committed when either a Type I algorithm was applied to solve a Type II problem, or a Type II algorithm was applied to solve a Type I problem. The means and standard deviations for the "errors of misclassification" are indicated in Table 4.

The multivariate analysis of variance indicated that significant differences existed between the groups on the number of "errors of misclassification" for the problems on Test AB and Test BA. The existence of large F-ratios for the multivariate test justified subsequent
### TABLE 3
SMALL GROUP AVERAGES FOR TEST AB AND TEST BA

<table>
<thead>
<tr>
<th>Treatment-Class</th>
<th>Test AB</th>
<th>Test BA</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB - I</td>
<td>5.50</td>
<td>1.50</td>
</tr>
<tr>
<td>AB - II</td>
<td>5.33</td>
<td>1.83</td>
</tr>
<tr>
<td>AB - III</td>
<td>4.83</td>
<td>1.17</td>
</tr>
<tr>
<td>AB - IV</td>
<td>5.33</td>
<td>1.00</td>
</tr>
<tr>
<td>BA - I</td>
<td>3.33</td>
<td>5.00</td>
</tr>
<tr>
<td>BA - II</td>
<td>1.66</td>
<td>5.17</td>
</tr>
<tr>
<td>BA - III</td>
<td>1.33</td>
<td>5.67</td>
</tr>
<tr>
<td>BA - IV</td>
<td>2.50</td>
<td>4.83</td>
</tr>
<tr>
<td>Control - I</td>
<td>3.83</td>
<td>2.33</td>
</tr>
<tr>
<td>Control - II</td>
<td>2.50</td>
<td>1.83</td>
</tr>
<tr>
<td>Control - III</td>
<td>1.83</td>
<td>1.50</td>
</tr>
<tr>
<td>Control - IV</td>
<td>2.83</td>
<td>1.50</td>
</tr>
</tbody>
</table>

### TABLE 4
MEANS AND STANDARD DEVIATIONS FOR ERRORS OF MISCLASSIFICATION

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Test AB Mean</th>
<th>Test AB SD</th>
<th>Test BA Mean</th>
<th>Test BA SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>.167</td>
<td>.236</td>
<td>4.000</td>
<td>.408</td>
</tr>
<tr>
<td>BA</td>
<td>3.125</td>
<td>.498</td>
<td>.250</td>
<td>.214</td>
</tr>
<tr>
<td>Control</td>
<td>.500</td>
<td>.360</td>
<td>1.000</td>
<td>.577</td>
</tr>
</tbody>
</table>
applications of univariate tests for each of the individual variables. The results are reported in Table 5.

**TABLE 5**

MULTIVARIATE AND UNIVARIATE ANALYSIS OF VARIANCE ERRORS OF MISCLASSIFICATION FOR VARIABLES TEST AB, TEST BA USING WILKS LAMBDA CRITERION

<table>
<thead>
<tr>
<th>Variables</th>
<th>Test</th>
<th>df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test AB, Test BA</td>
<td>M</td>
<td>(4,16)</td>
<td>50.842</td>
<td>.001</td>
</tr>
<tr>
<td>Test AB</td>
<td>U</td>
<td>(2,9)</td>
<td>72.695</td>
<td>.001</td>
</tr>
<tr>
<td>Test BA</td>
<td>U</td>
<td>(2,9)</td>
<td>86.549</td>
<td>.001</td>
</tr>
</tbody>
</table>

The analysis of variance described in Table 5 indicated that the treatment groups differed on the "errors of misclassification" for the two subtests. Plots of the cell means for the "errors of misclassification" are presented in Figure 6.

Any pairwise differences between the groups are to be tested. Tukey's test was applied to determine which differences in the cell means were significant. For Test AB a difference of .751 (p < .05) or .978 (p < .01) between two treatment means was necessary to conclude that the groups were different. Thus, the difference between the means of 2.958 favoring Treatment AB over Treatment BA was significant. The Control also committed fewer errors than Treatment BA on Test AB since the difference between the treatment means was 2.625. But the difference of
.333 between the Control and Treatment AB was not sufficient to conclude that the groups differed on Test AB at the .05 level.

On Test BA, a difference of .841 (p < .05) or 1.095 (p < .01) was necessary to conclude that two groups differed. Therefore, the difference of 3.750 between Treatment AB and Treatment BA was significant. In addition, the observed difference of 3.00 between the Control Treatment and Treatment AB was significant. However, the difference of .750 was not sufficient (p < .05) to conclude that the Control Treatment and Treatment BA differed on the errors of misclassification for Treatment BA.
The number of errors of misclassification for the individual small groups is presented in Table 6.

**TABLE 6**

**SMALL GROUP AVERAGES FOR ERRORS OF MISCLASSIFICATION**

<table>
<thead>
<tr>
<th>Treatment-Class</th>
<th>Test AB</th>
<th>Test BA</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB - I</td>
<td>.00</td>
<td>4.50</td>
</tr>
<tr>
<td>AB - II</td>
<td>.17</td>
<td>4.00</td>
</tr>
<tr>
<td>AB - III</td>
<td>.50</td>
<td>3.50</td>
</tr>
<tr>
<td>AB - IV</td>
<td>.00</td>
<td>4.00</td>
</tr>
<tr>
<td>BA - I</td>
<td>3.33</td>
<td>0.33</td>
</tr>
<tr>
<td>BA - II</td>
<td>3.00</td>
<td>0.50</td>
</tr>
<tr>
<td>BA - III</td>
<td>3.67</td>
<td>0.17</td>
</tr>
<tr>
<td>BA - IV</td>
<td>2.50</td>
<td>0.00</td>
</tr>
<tr>
<td>Control - I</td>
<td>0.83</td>
<td>1.50</td>
</tr>
<tr>
<td>Control - II</td>
<td>0.50</td>
<td>1.50</td>
</tr>
<tr>
<td>Control - III</td>
<td>0.00</td>
<td>0.50</td>
</tr>
<tr>
<td>Control - IV</td>
<td>0.67</td>
<td>0.50</td>
</tr>
</tbody>
</table>

**EXPERIMENT II**

**DESIGN AND HYPOTHESES**

Experiment II was a one-way design where the independent variable was the prior familiarity solving problems which were presented using a particular problem
form. Four dependent variables were measured for each of the subjects. These four variables were estimated by scores on four subtests. Each subtest consisted of problems which used a specific problem form. Figure 7 presents the design used for this experiment. Therefore, a multivariate analysis of variance was appropriate to analyze the results of this experiment.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Test A</th>
<th>Test B</th>
<th>Test C</th>
<th>Test D</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>W</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
</tr>
<tr>
<td>BBB</td>
<td>W</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
</tr>
<tr>
<td>ABC</td>
<td>W</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
</tr>
<tr>
<td>Control</td>
<td>W</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
</tr>
</tbody>
</table>

**FIGURE 7**

**DESIGN FOR EXPERIMENT II**

The hypotheses to be tested were as follows:

1. There were differences between the mean scores on Test A for the three instructional treatments.
2. There were differences between the mean scores on Test B for the three instructional treatments.
3. There were differences between the mean scores on Test C for the three instructional treatments.
4. There were differences between the mean scores on Test D for the three instructional treatments.
There were 213 subjects in the nine selected sixth-grade classrooms. A total of 203 subjects were assigned to one of 29 small groups. One person in each group was designated as the alternate and was to be excluded if there were no absences. Six subjects did not complete the experiment. Four subjects were replaced by the alternate group members because they missed the criterion test. Two members were replaced by alternate group members because of absentees during the instructional sessions. Some students who were absent were not excluded from the experiment. Five students were given individual sessions to "make-up" for the sessions they had missed due to absence. Three students were also given late exams due to their absences. One subject was excluded from the experiment due to his own request. Prior to the analysis, 23 alternates were excluded to obtain 29 small groups with six members each. Seven small groups received each of the three instructional treatments, and eight small groups were control groups. The analysis was applied to the small group means on the four subtests for the 21 small groups that participated in an instructional treatment. The eight control groups served only to allow qualitative comparisons and were not involved in the major hypotheses.

The statistical tests that were applied make the assumption of equal cell variances; but the tests remain
robust when the cell variances are not equal if the cell sizes are equal. The exclusion of the 23 subjects was not considered serious because there were 7 small groups receiving each of the instructional treatments. The unit of analysis was the small group and no small groups were excluded.

ANALYSIS

The criterion test consisted of four subtests. Each of the four subtests consisted of six items which used a specific problem form. Estimates of the four variables were obtained by using the subtest scores. These were denoted as the scores for Test A, Test B, Test C, and Test D. The unit of analysis was the small group. The individual means for each of the small groups on the four subtests can be found in Table 7. These means represent the average for the six members of each small group. To obtain a better understanding of these differences it was useful to look at the treatment means. The means and standard deviations for each treatment subtest combination are reported in Table 8. The means for the control groups are also reported. However, the analysis of variance was conducted after the control groups were excluded. The results for the control groups were obtained only to estimate how much the subjects improved during the instruction but were not
<table>
<thead>
<tr>
<th>Treatment-Class</th>
<th>Test A</th>
<th>Test B</th>
<th>Test C</th>
<th>Test D</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA - 1</td>
<td>5.67</td>
<td>4.17</td>
<td>3.67</td>
<td>4.33</td>
</tr>
<tr>
<td>AAA - 2</td>
<td>5.83</td>
<td>4.50</td>
<td>4.00</td>
<td>4.17</td>
</tr>
<tr>
<td>AAA - 4</td>
<td>5.17</td>
<td>3.33</td>
<td>2.50</td>
<td>3.00</td>
</tr>
<tr>
<td>AAA - 5</td>
<td>5.17</td>
<td>3.67</td>
<td>3.17</td>
<td>3.17</td>
</tr>
<tr>
<td>AAA - 6</td>
<td>5.00</td>
<td>3.67</td>
<td>3.33</td>
<td>3.17</td>
</tr>
<tr>
<td>AAA - 7</td>
<td>4.83</td>
<td>2.67</td>
<td>2.67</td>
<td>2.67</td>
</tr>
<tr>
<td>AAA - 8</td>
<td>5.17</td>
<td>3.33</td>
<td>2.50</td>
<td>2.67</td>
</tr>
<tr>
<td>BBB - 1</td>
<td>2.17</td>
<td>4.83</td>
<td>2.50</td>
<td>3.67</td>
</tr>
<tr>
<td>BBB - 3</td>
<td>3.67</td>
<td>5.17</td>
<td>3.67</td>
<td>4.33</td>
</tr>
<tr>
<td>BBB - 4</td>
<td>3.50</td>
<td>5.17</td>
<td>3.50</td>
<td>3.67</td>
</tr>
<tr>
<td>BBB - 5</td>
<td>4.00</td>
<td>5.67</td>
<td>4.00</td>
<td>4.33</td>
</tr>
<tr>
<td>BBB - 6</td>
<td>3.00</td>
<td>5.17</td>
<td>3.17</td>
<td>3.33</td>
</tr>
<tr>
<td>BBB - 7</td>
<td>2.83</td>
<td>4.67</td>
<td>3.50</td>
<td>3.33</td>
</tr>
<tr>
<td>BBB - 9</td>
<td>3.33</td>
<td>5.50</td>
<td>3.00</td>
<td>3.00</td>
</tr>
<tr>
<td>ABC - 2</td>
<td>4.17</td>
<td>4.50</td>
<td>4.33</td>
<td>4.50</td>
</tr>
<tr>
<td>ABC - 3</td>
<td>3.50</td>
<td>3.67</td>
<td>3.67</td>
<td>3.00</td>
</tr>
<tr>
<td>ABC - 4</td>
<td>4.17</td>
<td>4.50</td>
<td>4.50</td>
<td>3.50</td>
</tr>
<tr>
<td>ABC - 5</td>
<td>4.33</td>
<td>3.83</td>
<td>4.17</td>
<td>3.83</td>
</tr>
<tr>
<td>ABC - 6</td>
<td>3.50</td>
<td>3.83</td>
<td>3.83</td>
<td>3.33</td>
</tr>
<tr>
<td>ABC - 8</td>
<td>4.17</td>
<td>4.00</td>
<td>3.33</td>
<td>3.00</td>
</tr>
<tr>
<td>ABC - 9</td>
<td>4.00</td>
<td>4.33</td>
<td>4.17</td>
<td>3.00</td>
</tr>
<tr>
<td>Control - 1</td>
<td>1.67</td>
<td>2.83</td>
<td>2.50</td>
<td>2.67</td>
</tr>
<tr>
<td>Control - 2</td>
<td>2.00</td>
<td>2.50</td>
<td>2.33</td>
<td>2.50</td>
</tr>
<tr>
<td>Control - 3</td>
<td>1.50</td>
<td>2.17</td>
<td>2.33</td>
<td>2.83</td>
</tr>
<tr>
<td>Control - 5</td>
<td>2.17</td>
<td>1.50</td>
<td>1.50</td>
<td>2.00</td>
</tr>
<tr>
<td>Control - 6</td>
<td>2.00</td>
<td>2.00</td>
<td>2.00</td>
<td>2.17</td>
</tr>
<tr>
<td>Control - 7</td>
<td>2.00</td>
<td>2.83</td>
<td>1.67</td>
<td>1.83</td>
</tr>
<tr>
<td>Control - 8</td>
<td>2.17</td>
<td>2.17</td>
<td>1.67</td>
<td>2.50</td>
</tr>
<tr>
<td>Control - 9</td>
<td>1.83</td>
<td>2.83</td>
<td>2.50</td>
<td>2.17</td>
</tr>
</tbody>
</table>
involved in the major hypotheses.

TABLE 8

TREATMENT MEANS AND STANDARD DEVIATIONS FOR THE FOUR SUBTESTS

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Test A Mean</th>
<th>Test A SD</th>
<th>Test B Mean</th>
<th>Test B SD</th>
<th>Test C Mean</th>
<th>Test C SD</th>
<th>Test D Mean</th>
<th>Test D SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>5.263</td>
<td>.358</td>
<td>3.620</td>
<td>.599</td>
<td>3.120</td>
<td>.591</td>
<td>3.311</td>
<td>.675</td>
</tr>
<tr>
<td>BBB</td>
<td>3.214</td>
<td>.606</td>
<td>5.169</td>
<td>.343</td>
<td>3.384</td>
<td>.491</td>
<td>3.666</td>
<td>.508</td>
</tr>
<tr>
<td>Control</td>
<td>1.918</td>
<td>.236</td>
<td>2.354</td>
<td>.482</td>
<td>2.063</td>
<td>.409</td>
<td>2.396</td>
<td>.350</td>
</tr>
</tbody>
</table>

A multivariate analysis of variance was applied to the small group means to determine if differences existed between the three treatments on the four subtests. The KR - 20 reliability estimates were .89 for Test A, .81 for Test B, .73 for Test C, and .81 for Test D. Results on the multivariate analysis of variance are reported in Table 9. Because the multivariate analysis of variance indicated the existence of significant differences between the treatment groups, subsequent univariate tests were conducted. Results for each of the univariate tests are also reported in Table 9.

The results of the multivariate analysis of variance determined that differences existed between the treatments (p < .001). In addition, the univariate tests found significant differences at the .05 level on Tests A, B, and C.
TABLE 9

MULTIVARIATE AND UNIVARIATE ANALYSIS OF VARIANCE SCORES FOR THE FOUR SUBTESTS USING WILKS LAMBDA CRITERION

<table>
<thead>
<tr>
<th>Variables</th>
<th>Test</th>
<th>df</th>
<th>F</th>
<th>p &lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tests A, B, C, and D</td>
<td>M</td>
<td>(8,30)</td>
<td>31.577</td>
<td>.001</td>
</tr>
<tr>
<td>Test A</td>
<td>U</td>
<td>(2,18)</td>
<td>36.844</td>
<td>.001</td>
</tr>
<tr>
<td>Test B</td>
<td>U</td>
<td>(2,18)</td>
<td>22.092</td>
<td>.001</td>
</tr>
<tr>
<td>Test C</td>
<td>U</td>
<td>(2,18)</td>
<td>5.839</td>
<td>.011</td>
</tr>
<tr>
<td>Test D</td>
<td>U</td>
<td>(2,18)</td>
<td>0.652</td>
<td>.533</td>
</tr>
</tbody>
</table>

To determine the nature of the interactions between the treatments and the different subtest scores it is useful to plot the cell means. Plots of the cell means for Experiment II are presented in Figure 8.

FIGURE 8

PLOTS OF TREATMENT MEANS
Any pairwise differences between the cell means for the three treatments were of interest. The univariate tests indicated the existence of significant differences between the treatments for Tests A, B, and C at the .05 level. Tukey's Test was applied to determine any differences between the means for the three instructional treatments. Differences between treatment means of .616 on Test A, .609 on Test B, and .686 on Test C were significant (p < .05). Differences between the treatment means of .802 on Test A, .793 on Test B, and .893 on Test C were necessary to conclude the treatments differed (p < .01). All three pairwise comparisons for Test A were in excess of .616 and therefore significant. The difference between Treatment AAA and Treatment BBB was 2.049 favoring Treatment AAA. A difference of 1.286 was observed favoring Treatment AAA over Treatment ABC. A difference of .763 was sufficient to conclude that Treatment ABC was better (p < .05) than Treatment BBB on Test A.

Tukey's Test also determined that two of the three pairwise comparisons between the treatment groups for Test B were significant. A difference (p < .01) in favor of Treatment BBB of 1.549 was observed between Treatment AAA and Treatment BBB. The observed difference of 1.072 favoring Treatment BBB over Treatment ABC was also significant (p < .01). The difference between Treatment
AAA and Treatment ABC was equal to \(0.474 < 0.609\), and therefore, not significant \((p < 0.05)\).

Tukey's Test determined that there was one significant comparison for the treatment means on Test C. A difference of 0.686 was required to conclude that two means differed \((p < 0.05)\). Therefore, the difference of 0.880 favoring Treatment ABC over Treatment AAA was significant. But the difference of 0.666 favoring Treatment ABC over Treatment BBB was not significant. The difference between Treatment AAA and Treatment BBB was not significant on Test C. A summary of the significant comparisons between treatments has been reported in Table 10.

**TABLE 10**

MEANS AND SIGNIFICANT COMPARISONS FOR THE FOUR SUBTESTS

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Test A</th>
<th>Test B</th>
<th>Test C</th>
<th>Test D</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>5.262</td>
<td>3.620</td>
<td>3.120</td>
<td>3.311</td>
</tr>
<tr>
<td>BBB</td>
<td>3.214</td>
<td><strong>5.169</strong></td>
<td><strong>3.334</strong></td>
<td><em>3.366</em></td>
</tr>
<tr>
<td>ABC</td>
<td><em>3.977</em></td>
<td>4.097</td>
<td><strong>4.000</strong></td>
<td>3.451</td>
</tr>
<tr>
<td>CONTROL</td>
<td>1.918</td>
<td>2.354</td>
<td>2.063</td>
<td>2.396</td>
</tr>
</tbody>
</table>

\(\ast p < 0.05\)

\(\ast\ast p < 0.01\)

The univariate test for Test D did not yield a significant difference at the .05 level. Therefore, no tests were applied to detect differences between the
individual treatments.

HIGH AND LOW ABILITY

Krutetskii (1976) through the use of case studies has determined that there were differences between the way that high and low ability students remember problems that they have previously solved. High ability students tend to remember the mathematical structure of problems; but low ability students recall the contextual details in the problems.

The Mathematical Applications Section of the Stanford Achievement Test (1971) was given to each of the students prior to the treatments. This test was used to help stratify the students into high and low ability. The subjects were then assigned at random to one of the three instructional treatments or the control treatment so that there were three high and three low ability students in each group.

An analysis of variance was not applied to the results obtained for the high and low ability groups because of the violation of several assumptions.

1. The small groups of high ability students were not instructed separately from the low ability students.

2. The criterion for high ability varied between the different classes. This ranged from 27 to 31 for the different classes.
3. If the criterion for high ability were made uniform, the small groups would not have equal numbers of subjects.

Although an analysis of variance was not applied, the results based on ability level are of interest. The median for all subjects participating in the study was 30. Thus, it is reasonable to define high and low ability students as follows:

High Ability - A subject is of high ability if his score is greater than or equal to 30.

Low Ability - A subject is of low ability if his score is less than 30.

Based on these definitions the means for high and low ability students are designated in Table 11 and Table 12.

SUMMARY

Significant differences were observed between the treatment groups on both Experiment I subtests. Prior familiarity with certain problem type-form combinations was found to be a powerful factor when these type-form combinations were not systematically varied during the instruction. Treatment AB subjects scored higher on

*Scores for the Mathematics Applications section for the Stanford Achievement Test Intermediate Level II Form A.
TABLE 11

TREATMENT MEANS FOR HIGH ABILITY SUBJECTS

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Test A</th>
<th>Test B</th>
<th>Test C</th>
<th>Test D</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>5.83</td>
<td>4.25</td>
<td>4.23</td>
<td>4.18</td>
</tr>
<tr>
<td>n = 24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BBB</td>
<td>4.29</td>
<td>5.74</td>
<td>4.05</td>
<td>4.82</td>
</tr>
<tr>
<td>n = 24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABC</td>
<td>4.95</td>
<td>4.95</td>
<td>5.22</td>
<td>4.67</td>
</tr>
<tr>
<td>n = 22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>2.73</td>
<td>3.50</td>
<td>2.95</td>
<td>3.85</td>
</tr>
<tr>
<td>n = 22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Test AB than either the Control or Treatment BA subjects.

TABLE 12

TREATMENT MEANS FOR LOW ABILITY SUBJECTS

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Test A</th>
<th>Test B</th>
<th>Test C</th>
<th>Test D</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>4.44</td>
<td>2.72</td>
<td>1.64</td>
<td>2.00</td>
</tr>
<tr>
<td>n = 18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BBB</td>
<td>2.31</td>
<td>4.24</td>
<td>2.31</td>
<td>2.25</td>
</tr>
<tr>
<td>n = 18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABC</td>
<td>2.38</td>
<td>2.80</td>
<td>2.69</td>
<td>2.00</td>
</tr>
<tr>
<td>n = 20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>1.13</td>
<td>1.45</td>
<td>1.11</td>
<td>1.11</td>
</tr>
<tr>
<td>n = 20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

On the other hand, Treatment BA subjects scored higher on Test BA than either the Control or Treatment AB small groups.
The results for Experiment I also supported the hypothesis that presenting sets of instances which were convergent and not matched, would result in many students forming misconceptions. As predicted, Treatment AB subjects committed more "errors of misclassifications" on Test BA than did either of the other treatments. On Test AB the Treatment BA subjects committed more "errors of misclassification" than did either Treatment AB or the Control.

In Experiment II, the hypothesis that prior familiarity with a problem form would result in increased achievement on problems using that problem form, was supported over and over. Treatment AAA small groups scored higher \( (p < .01) \) on Form A problems than either Treatment BBB or Treatment ABC small groups. Treatment BBB small groups scored higher \( (p < .01) \) than the Treatment AAA or Treatment ABC small groups on Form B problems. Treatment ABC subjects scored higher \( (p < .05) \) on Form A problems than Treatment BBB subjects. Treatment ABC subjects also scored higher \( (p < .05) \) on Test C than the Treatment AAA subjects. Two comparisons were not significant \( (p < .05) \), but the differences between the means favored Treatment ABC. These were: Treatment ABC compared with Treatment AAA on Form B problems, and Treatment ABC compared with Treatment BBB on Form C problems.
The major hypothesis that varying the contextual details and problem forms used to construct the sets of exercises would result in higher achievement on previously unencountered forms was not supported. According to Markle (1970), it is reasonable to predict that Treatment ABC would score the highest on Form D, the form which was not encountered by any of the subjects during the instructional treatments. However, no differences were observed (p < .533) between the instructional treatments on Test D.
CHAPTER V

CONCLUSIONS

Both experiments resulted in significant differences among the treatment groups. To better understand these differences, it is important to consider the relationship between the two experiments. The treatments for the first experiment provided practice in solving two problem type-form combinations. There were two problem type-form combinations presented in each treatment. These combinations were convergent and not matched. According to Tennyson (1973), this would result in "errors of misclassification."

The treatments for the second experiment provided practice in solving three problem types. The treatments were determined by the problem form used to present the exercises. In two of the three instructional treatments, the problem form was never varied. In one treatment, the problem form was varied on each of the three days. The problems that were presented in the two treatments which did not change the problem form, can be described as convergent and matched. Merrill (1971) has predicted that such a set of exercises would result in the subjects
committing errors of undergeneralization.

EXPERIMENT I

There were differences \((p < .01)\) between the instructional treatment groups on both Test AB and Test BA. As indicated previously, the items on the two subtests were matched to items which were presented during the two treatments.

Test AB consisted of six items which used the two problem type-form combinations encountered by Treatment AB subjects. Test BA consisted of six items which employed the two problem type-form combinations encountered by Treatment BA subjects. There were four small groups that received no instruction on solving the word problems. These groups were designated as control groups and enabled comparisons that would determine if the instruction might even lower the level of achievement on specific items. A summary of the subtest scores are presented in Table 13.

A multivariate analysis of variance determined that significant differences existed \((p < .001)\) on the two subtests scores. Subsequent univariate tests also determined that differences existed \((p < .001)\) for both subtests. Applications of Tukey's Test for Honest Significant Differences determined that differences existed between the treatments. From Table 13 it was
clear that the scores on Test AB for Treatment AB were significantly higher than the scores for either Treatment BA or the Control Treatment. But the difference favoring the Control Treatment over Treatment BA was not significant. However, on Test BA these results were reversed with the scores for Treatment BA significantly higher than the scores for either Treatment AB or the Control. Again the difference which favored the Control over Treatment AB was not significant.

The results for Test AB and Test BA were also analyzed according to the "errors of misclassification." Table 14 presents a summary of these results. Univariate tests determined that significant differences existed (p < .001) between the three treatments on the "errors of misclassification" for both Test AB and Test BA. Treatment BA small groups committed more "errors of misclassification" (p < .01) on Test AB than either Treatment AB or the Control small groups. The difference
between the Control Treatment and Treatment AB for Test AB was not significant. The difference between the Control Treatment and Treatment AB was not significant for Test AB. On Test BA, Treatment AB small groups committed more "errors of misclassification" (p < .01) than either Treatment BA or the Control. Again the difference between the Control Treatment and Treatment BA was not significant.

To interpret these results it is important to recall some features of each treatment. Treatment AB presented only Type I - Form A and Type II - Form B problems. Treatment BA presented only Type I - Form B and Type II - Form A problems. According to Merrill (1973) and Tennyson (1972b, 1974b), when the irrelevant attributes are not varied systematically, these attributes become associated with the relevant attributes and the method for solution. The data for the "errors of
misclassification" supported the conjecture that the contextual details and the problem form become strongly associated with the algorithm.

There were two comparisons which displayed an interference. This interference could occur when students encountered a problem that employed a problem type-form combination not matched with the combinations encountered during instruction. These two comparisons are indicated by the following:

(1) Treatment AB compared with Treatment BA on Test AB, and

(2) Treatment BA compared with Treatment AB on Test BA.

When the problem type-form combinations were matched with those problems encountered during instruction, the subjects performed well. However, when the problem type-form combinations were reversed and did not match the combinations encountered during instruction, the performance declined markedly. Moreover, as indicated by the "errors of misclassification," the subjects did apply the incorrect algorithm to roughly 50% of those problems which did not match the problem type-form combinations encountered during instruction. The Control Treatment small groups appeared to score higher than Treatment AB on Test BA. The control small groups also appeared to score higher than Treatment BA on Test AB. However, these comparisons were not significant (p < .05). If
these differences had been significant, this would indicate that there was a drop in performance "caused" by the instruction. In terms of the total score, this difference was not significant. But if one analyzes the "errors of misclassification," there was an increase in the number of errors of a specific type for the subtests where the problem type-form combinations were not matched to the problems encountered during instruction.

In a sense, there was a contradiction when the instructional treatments were compared to the Control Treatment on those problems which did not match the form-type combinations encountered during instruction. On the one hand, the differences between the treatments on the subtest scores were not significant. On the other hand, there was an increase in the number of errors of a specific type for the subjects who received the instructional treatments. There seem to be two explanations for this contradiction. First, prior exposure to the form enabled some subjects to gain a better understanding of the relationships which existed in these problems. This enabled some subjects to perform better on the items. Thus, the positive effect of familiarity with the problem form cancelled out the negative effect of interference. Therefore, no differences between the control and the instructional groups were observed. The other explanation, is also plausible. The Control Treatment performed
better than the instructional treatments on the subtest which used problems not matched to those encountered during the instructional sessions; but the design was not sensitive enough to detect this difference.

EXPERIMENT II

Significant differences between the treatment means were found for three of the four subtests. As detailed in Chapter III, each subtest was derived using a specific problem form. Three problem forms were encountered during the treatments by some of the students. Some subjects encountered the same problem form during each of their three instructional sessions. Other groups solved problems of a specific problem form for only one session. Still others did not encounter problems of a specific problem form during any sessions. The means for each treatment are presented in Table 15. Significant differences between the means are also indicated.

It is useful to interpret the results by organizing the comparisons into four categories. First, there were comparisons between treatments when subjects encountered a problem form during all three instructional sessions and treatments where the subjects did not encounter the problem form. Secondly, there were two comparisons between treatments where subjects had encountered a problem form three times and a treatment where subjects
TABLE 15
TREATMENT MEANS AND SIGNIFICANT COMPARISONS FOR THE FOUR SUBTESTS

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Test A</th>
<th>Test B</th>
<th>Test C</th>
<th>Test D</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>5.263</td>
<td>3.620</td>
<td>3.120</td>
<td>3.311</td>
</tr>
<tr>
<td>BBB</td>
<td>3.214</td>
<td>5.169</td>
<td>3.334</td>
<td>3.666</td>
</tr>
<tr>
<td>ABC</td>
<td>3.977</td>
<td>4.097</td>
<td>4.000</td>
<td>3.451</td>
</tr>
<tr>
<td>Control</td>
<td>1.918</td>
<td>2.354</td>
<td>2.063</td>
<td>2.396</td>
</tr>
</tbody>
</table>

*p < .05
**p < .01

had encountered the problem form for only one session.

Thirdly, there were four comparisons between a treatment where the subjects encountered a problem form during one session and a treatment where the subjects had not encountered the problem form. Finally, there was the comparison between subjects for the three instructional treatments on a problem form which had not been encountered by any of the groups during instruction. Table 16 has organized the means by the number of sessions spent solving the different problem forms.

The two comparisons between groups that encountered a problem form three times and groups which had not encountered the problem form were significant. The difference between Treatment AAA and Treatment BBB on Test A was significant \(p < .01\) in favor of Treatment AAA. Comparing the same two groups on Test B the
difference was also significant \( (p < .01) \); but this difference as expected, favored the Treatment BBB groups.

**TABLE 16**

RESULTS WHEN GROUPED BY THE NUMBER OF SESSIONS USING EACH PROBLEM FORM

<table>
<thead>
<tr>
<th>Sessions Using Form</th>
<th>Treatment</th>
<th>Test</th>
<th>Mean</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 A A A A</td>
<td>AAA</td>
<td>A</td>
<td>5.263</td>
<td>87.7</td>
</tr>
<tr>
<td>3 BBB B</td>
<td>BBB</td>
<td>B</td>
<td>5.169</td>
<td>86.2</td>
</tr>
<tr>
<td>3 1 ABC A</td>
<td>ABC</td>
<td>A</td>
<td>3.977</td>
<td>66.1</td>
</tr>
<tr>
<td>3 1 ABC B</td>
<td>ABC</td>
<td>B</td>
<td>4.094</td>
<td>68.2</td>
</tr>
<tr>
<td>3 1 ABC C</td>
<td>ABC</td>
<td>C</td>
<td>4.000</td>
<td>66.7</td>
</tr>
<tr>
<td>3 0 AAA B</td>
<td>AAA</td>
<td>B</td>
<td>3.620</td>
<td>60.3</td>
</tr>
<tr>
<td>3 0 AAA C</td>
<td>AAA</td>
<td>C</td>
<td>3.120</td>
<td>52.0</td>
</tr>
<tr>
<td>3 0 BBB A</td>
<td>BBB</td>
<td>A</td>
<td>3.214</td>
<td>53.6</td>
</tr>
<tr>
<td>3 0 BBB C</td>
<td>BBB</td>
<td>C</td>
<td>3.334</td>
<td>55.6</td>
</tr>
<tr>
<td>3 0 AAA D</td>
<td>AAA</td>
<td>D</td>
<td>3.311</td>
<td>55.2</td>
</tr>
<tr>
<td>3 0 BBB D</td>
<td>BBB</td>
<td>D</td>
<td>3.666</td>
<td>61.1</td>
</tr>
<tr>
<td>3 0 ABC D</td>
<td>ABC</td>
<td>D</td>
<td>3.451</td>
<td>57.5</td>
</tr>
<tr>
<td>0 0 Control A</td>
<td>Control</td>
<td>A</td>
<td>1.917</td>
<td>32.0</td>
</tr>
<tr>
<td>0 0 Control B</td>
<td>Control</td>
<td>B</td>
<td>2.354</td>
<td>39.2</td>
</tr>
<tr>
<td>0 0 Control C</td>
<td>Control</td>
<td>C</td>
<td>2.063</td>
<td>34.4</td>
</tr>
<tr>
<td>0 0 Control D</td>
<td>Control</td>
<td>D</td>
<td>2.396</td>
<td>39.9</td>
</tr>
</tbody>
</table>

Significant differences were found both times when the treatment groups which encountered a problem form for three sessions were compared with the treatment groups which encountered the problem form for only one session. A significant difference \( (p < .01) \) was found favoring Treatment AAA over Treatment ABC on Test A. A difference \( (p < .01) \) was also found favoring Treatment BBB over Treatment ABC on Test B.
The differences which result from comparing treatments which encounter a problem form once, with treatments that did not encounter the problem form were not conclusive. There were four such comparisons. All four of the differences tended to favor the groups which had encountered the problem form. However, only two of these differences were statistically significant \((p < .05)\). The two comparisons which were significant were: (1) the difference favoring Treatment ABC over Treatment BBB on Test C, and (2) the difference favoring Treatment ABC over Treatment AAA on Test C. The remaining comparisons were not statistically significant. These were: (1) a difference favoring Treatment ABC over Treatment AAA on Test B, and (2) a difference favoring Treatment ABC over Treatment BBB on Test C.

The number of days that subjects had encountered the problem forms seemed to be a consistent indicator of their performance on selected subtests. There were two treatments where subjects encountered a problem form during three sessions. Treatment AAA subjects averaged 87.7% on Form A problems and Treatment BBB subjects averaged 86.1% on the Form B problems. Treatment ABC subjects encountered three problem forms during the instruction. They had solved problems using each of these forms during one instructional session. The average performance was consistent on the three problem forms. These averages ranged from 66.3% to 68.2% correct.
Treatment AAA subjects did not encounter three of the problem forms. Treatment BBB subjects also did not encounter three of the problem forms. Treatment ABC subjects did not encounter Form D. The averages for these seven cells did not show a marked variability ranging from 52% to 61%.

Although the control groups were not included in the analysis of variance, their means were reported in Table 16. The control groups scored lower than any instructional treatment groups on all of the problem forms. The groups that did not receive any instruction averaged about 36% on the different problem forms. Therefore, there was evidence that the students that participated in the instructional sessions did transfer skills to solve problems with an unencountered problem form.

In addition to the six significant comparisons presented earlier which indicated that prior familiarity with a problem form tended to reduce the difficulty of similar problems, the qualitative summary presented above supports this conjecture. In fact, it provides very strong support because of the consistency of the observed differences between the groups.

A second conjecture consistent with Merrill (1971), Markle (1970), and Tennyson (1973), was that when the problem form was varied during instruction, this will result in higher performance on items with a previously unencountered form. This conjecture would predict
that Treatment ABC should score higher on Test D than the other two treatments. However, the results of the univariate test did not find differences ($p < .533$) between the groups. In addition, Table 15 indicates that the average for the Treatment ABC groups was not higher than the other treatment groups on Test D.

The data supports the conjecture that prior familiarity with a given problem form and type combination reduces the difficulty level for these items. The main effects for Test A, Test B, and Test C support this conjecture not once but several times. For instance, Treatment AAA was superior to Treatment BBB on Form A, and at the same time Treatment BBB subjects were superior to Treatment AAA subjects on Form B.

The groups were constructed using stratified random sampling to assure that they were equal in ability prior to the treatments. Therefore, there does not appear to be a plausible explanation for the observed differences other than the treatments. The differences between the treatments were determined by the specific problem forms employed to present the items. Despite the fact that the differences can be attributed to the treatments, several questions remain. Since several irrelevant properties were varied to derive the problem forms, it was unclear what part individual changes in problems had on the students' abilities to solve the problems.
When two problems of different forms are given to a subject, there is not a simple explanation which accounts for a subject being able to solve only the problem which matched the problem form-type combination encountered during instruction. Many factors may contribute to reducing the difficulty of problems which have a problem type-form combination that has previously been encountered. Some of the more important factors are: (a) the student's reading ability, (b) comprehension of the relationship between the variables, (c) similarity of the position of the numbers in the problem statement, and (d) similarity of the relative size of the numbers used in the problem statement. It seems likely that the observed differences between the respective treatments can be attributed to the problem forms encountered during the instructional sessions; but this difference is due to the cumulative effect of changing several of the characteristics used to derive the different problem forms. Further research is necessary to determine the effect of changing smaller subsets of the irrelevant attributes.

No differences were observed between the treatments on Test D, which represented a problem form which was not encountered by any of the subjects during the treatments. According to Markle (1969), Treatment ABC might have been expected to score higher because the problem form was varied during instruction. Therefore, this research does
not support the teaching theory that varying the irrelevant attributes during instruction will result in higher achievement on previously unencountered instances. However, it would be premature to conclude it is not helpful to systematically vary irrelevant properties during instruction. First, the instructional sessions for the subjects lasted less than 90 minutes. The likelihood that any treatment of less than 90 minutes will result in a significant change in a subject's ability to transfer knowledge is very small. To significantly improve a subject's ability to transfer knowledge to unencountered instances is a long term objective of many educators and is not likely to be obtained easily. Second, the specific mathematics problems which were selected for this research may not have been sensitive for measuring transfer to unencountered forms. Third, in Treatment ABC the problem form was varied in the sense that different problem forms were used each day; but variability was not introduced within each instructional session. Finally, Krutetskii (1976), predicted that high and low ability students are affected by variability in different ways. Good students observe problems as similar based on their structure and are able to transfer knowledge easily when the problems "contextual details" have been changed. However, the poor student treats each problem which uses different contextual details as a
different problem type. Thus, he may become so totally confused that he is unable to solve any problems without help. Further research is necessary to determine if high and low ability students react differently to variation among the irrelevant properties when presenting a set of examples. If this difference does exist, it would in part explain why no differences were observed between the treatments on Test D. The results for the high ability and low ability students were presented in Table 10 and Table 11. There did not appear to be any major interactions which were different for high and low ability students. The only difference appears to be that the high ability students scored uniformly higher than the low ability subjects.

Finally, the results may have been affected because Test C and D were given one or two days after Tests A and B. This introduces the possibility that the subjects either learned solution techniques by simply taking Test A and B, or the subjects may have forgotten some of the methods used for solving these problems between the two testings. Additional research should examine retention after two or three weeks.

SUMMARY AND RECOMMENDATIONS

The analysis supported the hypothesis that familiarity with a specific problem type-form combination
reduces the difficulty of subsequent problems which employed the same problem type-form combination. It was also clear that when the problem form was not systematically varied during instruction, the problem form frequently became associated with the solution algorithm. The hypothesis that higher variability among the irrelevant properties during instruction results in improved performance on problems using unencountered contextual detail, was not substantiated. Rather questions were raised regarding the construction of an optimal set of sample problems designed to produce transfer.

Prior familiarity with a specific problem form affected the ability of many subjects to work the problems. The differences between performance on selected problem forms not only consistently favored the subjects with prior familiarity with the particular type-form combination, but the differences were large. The problem forms differed only on contextual details and other relatively unimportant characteristics. The algorithms which were used to solve the problems were identical. One could easily argue that solving two problems of the same type does not require transfer, or requires only a minimal transfer. But it was observed that many students were unable to make this minimal transfer. As noted in Table 15, students averaged over 85% for those problems that used the problem that they had
encountered for three sessions, but less than 60% on those items with a previously unencountered form.

Despite the fact that this experiment did not provide an optimal strategy which can be used to generate a set of problems, there are several implications which may be applied to the classroom. Perhaps the most immediate application is in the area of test construction. This research has determined a difference in difficulty for problems which contained only number changes in previously encountered problems, and problems which contained number changes, changes in the contextual details, and other irrelevant properties. Therefore, to construct an item designed to measure transfer, more than the numbers must be changed. This means that an item which requires complex cognitive skill for one group of students, may not require the same cognitive level for an equivalent group of students which have encountered a similar item. Therefore, if students have access to prior examinations in which a problem may have required a high cognitive level skill to solve, it may not require the same skill on a second testing.

There are applications for designing exercises. Mass practice when the contextual details are not varied must be avoided. The use of behavioral objectives which are so specific that only number changes in "sample test"
items are used to form test items, must be avoided. If not, only a memorizational level may be measured.

The fact that the difficulty of a problem can be significantly reduced by providing the student with sample “test items” which vary from the criterion items by only the numbers in the problem statement, must be considered when analyzing the results for specific studies. Some studies provide students with "sample tests" and report substantial improvement due to some treatment. To interpret these results, the effect of reducing the difficulty of the criterion items by providing the sample items must be considered as an alternate explanation for the improvement.

Further research is needed to answer several questions. Specific irrelevant properties such as the topic, contextual details, order of the data, and use of key words, were varied to derive the different problem forms. Research is needed to determine the relative importance of each of these factors. High and low ability blocking should be investigated to determine if differences in the effect of problem form exist for different ability levels. Even though varying the problem form was not found to be helpful in solving previously unencountered problems, research is needed to investigate if there are long term effects due to variability of irrelevant properties. This study involved specific sets of word
problems at the fifth and sixth grade level. There is a need to look at different kinds of problems and different grade levels. Detailed case studies are needed to determine why some students generalize to unencountered instances while others do not. This study has verified that the problem form under specific conditions can produce a powerful effect on a subject's ability to solve similar problems. In this sense this research has made a contribution, but many questions have been raised. If these questions can be answered, this will result in a much deeper understanding of the importance of familiarity with problem form and contextual details in mathematical learning.
APPENDIX A

PROBLEMS SOLVED DURING TREATMENTS
FOR EXPERIMENT I

Copies of the problems solved during the two instructional treatments are presented on the following twelve pages. The first six pages are the problems presented to Treatment AB subjects. The next six pages are the problems presented to Treatment BA subjects. The pages are in the same order that they were presented to the subjects. Two pages were presented during the first instructional session; the remaining four pages were presented during the second session.
Two men left Columbus in a car on a long trip. One man drove at 35 miles per hour for 6 hours. The other man drove at 45 miles per hour for 10 hours to complete the trip. How far was the trip?

Two men left Columbus in a car on a long trip. One man drove at 42 miles per hour for 4 hours. The other man drove at 30 miles per hour for 8 hours to complete the trip. How far was the trip?

Two men left Columbus in a car on a long trip. One man drove at 40 miles per hour for 12 hours. The other man drove at 60 miles per hour for 7 hours to complete the trip. How far was the trip?
A man accepted a job where he was paid $7 each day that he was training. His pay was then increased to $11 for each day after his training period had finished. His training period lasted for 5 days. Including his training period, he worked 14 days. How much did he earn on this job?

A man accepted a job where he was paid $10 each day that he was training. His pay was then increased to $15 for each day after his training period had finished. His training period lasted for 4 days. Including his training period, he worked 11 days. How much did he earn on this job?

A man accepted a job where he was paid $11 each day that he was training. His pay was then increased to $15 for each day after his training period had finished. His training period lasted for 9 days. Including his training period, he worked 20 days. How much did he earn on this job?
Two men left Columbus in a car on a long trip. One man drove at 36 miles per hour for 5 hours. The other man drove at 50 miles per hour for 7 hours to complete the trip. How far was the trip?

Two men left Columbus in a car on a long trip. One man drove at 50 miles per hour for 8 hours. The other man drove at 60 miles per hour for 12 hours to complete the trip. How far was the trip?

Two men left Columbus in a car on a long trip. One man drove at 55 miles per hour for 10 hours. The other man drove at 40 miles per hour for 6 hours to complete the trip. How far was the trip?
A man accepted a job where he was paid $5 each day that he was training. His pay was then increased to $8 for each day after his training period had finished. His training period lasted for 4 days. Including his training period, he worked 13 days. How much did he earn on this job?

A man accepted a job where he was paid $7 each day that he was training. His pay was then increased to $10 for each day after his training period had finished. His training period lasted for 8 days. Including his training period, he worked 20 days. How much did he earn on this job?

A man accepted a job where he was paid $8 each day that he was training. His pay was then increased to $12 for each day after his training period had finished. His training period lasted for 10 days. Including his training period, he worked 25 days. How much did he earn on this job?
Two men left Columbus in a car on a long trip. One man drove at 36 miles per hour for 14 hours. The other man drove at 52 miles per hour for 11 hours to complete the trip. How far was the trip?

Two men left Columbus in a car on a long trip. One man drove at 50 miles per hour for 11 hours. The other man drove at 55 miles per hour for 23 hours to complete the trip. How far was the trip?

Two men left Columbus in a car on a long trip. One man drove at 55 miles per hour for 12 hours. The other man drove at 40 miles per hour for 27 hours to complete the trip. How far was the trip?
A man accepted a job where he was paid $15 each day that he was training. His pay was then increased to $25 for each day after his training period had finished. His training period lasted for 14 days. Including his training period, he worked 60 days. How much did he earn on this job?

A man accepted a job where he was paid $18 each day that he was training. His pay was then increased to $24 for each day after his training period had finished. His training period lasted for 14 days. Including his training period, he worked 45 days. How much did he earn on this job?

A man accepted a job where he was paid $20 each day that he was training. His pay was then increased to $28 for each day after his training period had finished. His training period lasted for 30 days. Including his training period, he worked 60 days. How much did he earn on this job?
Two men left Columbus in a car on a long trip. One man drove at 35 miles per hour for 7 hours. The other man drove at 40 miles per hour so that the trip was completed in 16 hours. How far was the trip?

Two men left Columbus in a car on a long trip. One man drove at 60 miles per hour for 4 hours. The other man drove at 44 miles per hour so that the trip was completed in 11 hours. How far was the trip?

Two men left Columbus in a car on a long trip. One man drove at 50 miles per hour for 7 hours. The other man drove at 45 miles per hour so that the trip was completed in 15 hours. How far was the trip?
A man accepted a job where he was paid $9 each day that he was training. His pay was then increased to $11 for each day after his training period had finished. His training period lasted for 6 days. After his training period, he worked 30 days. How much did he earn on this job?

A man accepted a job where he was paid $8 each day that he was training. His pay was then increased to $10 for each day after his training period had finished. His training period lasted for 7 days. After his training period, he worked 8 days. How much did he earn on this job?

A man accepted a job where he was paid $10 each day that he was training. His pay was then increased to $13 for each day after his training period had finished. His training period lasted for 5 days. After his training period, he worked 12 days. How much did he earn on this job?
Two men left Columbus in a car on a long trip. One man drove at 41 miles per hour for 8 hours. The other man drove at 50 miles per hour so that the trip was completed in 15 hours. How far was the trip?

Two men left Columbus in a car on a long trip. One man drove at 35 miles per hour for 5 hours. The other man drove at 40 miles per hour so that the trip was completed in 12 hours. How far was the trip?

Two men left Columbus in a car on a long trip. One man drove at 60 miles per hour for 8 hours. The other man drove at 50 miles per hour so that the trip was completed in 14 hours. How far was the trip?
A man accepted a job where he was paid $6 each day that he was training. His pay was then increased to $9 for each day after his training period had finished. His training period lasted for 7 days. After his training period, he worked 20 days. How much did he earn on this job?

A man accepted a job where he was paid $9 for each day that he was training. His pay was then increased to $15 for each day after his training period had finished. His training period lasted for 5 days. After his training period, he worked 8 days. How much did he earn on this job?

A man accepted a job where he was paid $10 each day that he was training. His pay was then increased to $12 for each day after his training period had finished. His training period lasted for 8 days. After his training period, he worked 15 days. How much did he earn on this job?
Two men left Columbus in a car on a long trip. One man drove at 45 miles per hour for 12 hours. The other man drove at 53 miles per hour so that the trip was completed in 20 hours. How far was the trip?

Two men left Columbus in a car on a long trip. One man drove at 35 miles per hour for 11 hours. The other man drove at 44 miles per hour so that the trip was completed in 12 hours. How far was the trip?

Two men left Columbus in a car on a long trip. One man drove at 50 miles per hour for 12 hours. The other man drove at 40 miles per hour so that the trip was completed in 26 hours. How far was the trip?
A man accepted a job where he was paid $20 each day that he was training. His pay was then increased to $32 for each day after his training period had finished. His training period lasted for 24 days. After his training period, he worked 56 days. How much did he earn on this job?

A man accepted a job where he was paid $30 each day that he was training. His pay was then increased to $40 for each day after his training period had finished. His training period lasted for 16 days. After his training period, he worked 48 days. How much did he earn on this job?

A man accepted a job where he was paid $16 each day that he was training. His pay was then increased to $25 for each day after his training period had finished. His training period lasted for 21 days. After his training period, he worked 54 days. How much did he earn on this job?
APPENDIX B
TEST ITEMS FOR EXPERIMENT I

A copy of the test is provided on the next four pages. The first two pages constitute Test AB. The third and fourth pages constitute Test BA. The order of the pages was varied so that only one fourth of the subjects were given a test with the pages in this order.
Two men left Columbus in a car on a long trip. One man drove at 40 miles per hour for 7 hours. The other man drove at 50 miles per hour for 8 hours to complete the trip. How far was the trip?

Two men left Columbus in a car on a long trip. One man drove at 45 miles per hour for 6 hours. The other man drove at 55 miles per hour for 8 hours to complete the trip. How far was the trip?

Two men left Columbus in a car on a long trip. One man drove at 60 miles per hour for 4 hours. The other man drove at 42 miles per hour for 11 hours to complete the trip. How far was the trip?
A man accepted a job where he was paid $7 each day that he was training. His pay was then increased to $11 for each day after his training period had finished. His training period lasted for 6 days. Including his training period, he worked 16 days. How much did he earn on this job?

A man accepted a job where he was paid $6 each day that he was training. His pay was then increased to $12 for each day after his training period had finished. His training period lasted for 8 days. Including his training period, he worked 20 days. How much did he earn on this job?

A man accepted a job where he was paid $10 each day that he was training. His pay was then increased to $15 for each day after his training period had finished. His training period lasted for 12 days. Including his training period, he worked 25 days. How much did he earn on this job?
Two men left Columbus in a car on a long trip. One man drove at 35 miles per hour for 3 hours. The other man drove at 50 miles per hour so that the trip was completed in 10 hours. How far was the trip?

Two men left Columbus in a car on a long trip. One man drove at 40 miles per hour for 6 hours. The other man drove at 65 miles per hour so that the trip was completed in 9 hours. How far was the trip?

Two men left Columbus in a car on a long trip. One man drove at 60 miles per hour for 8 hours. The other man drove at 42 miles per hour so that the trip was completed in 20 hours. How far was the trip?
A man accepted a job where he was paid $9 each day that he was training. His pay was then increased to $12 for each day after his training period had finished. His training period lasted for 5 days. After his training period, he worked 10 days. How much did he earn on this job?

A man accepted a job where he was paid $7 each day that he was training. His pay was then increased to $15 for each day after his training period had finished. His training period lasted for 6 days. After his training period, he worked 18 days. How much did he earn on this job?

A man accepted a job where he was paid $8 each day that he was training. His pay was then increased to $10 for each day after his training period had finished. His training period lasted for 8 days. After his training period, he worked 12 days. How much did he earn on this job?
APPENDIX C

PROBLEMS SOLVED DURING TREATMENTS
FOR EXPERIMENT II

Copies of the problems solved during the three instructional treatments are presented on the next eighteen pages. The first six pages list the problems presented to Treatment AAA small groups. The next six pages list the problems presented to Treatment BBB subjects, and the problems presented to the Treatment ABC subjects are listed on the last six pages. Two pages were presented to the subjects during each of the three instructional sessions. The pages are in the same order that they were presented to the subjects during the treatments.
Two men left Columbus in a car on a 360 mile trip. One man drove at 44 miles per hour for 6 hours. What was the average speed of the other driver if the trip was completed in 9 hours?

Two men left Columbus in a car on a long trip. One man drove at 60 miles per hour for 15 hours. The other man averaged 45 miles an hour for 10 hours to complete the trip. How far was the trip?

Two men left Columbus in a car on a long trip. One man drove at 32 miles per hour for 4 hours. The other man drove 52 miles an hour for 6 hours to complete the trip. What was the average speed for the trip?
Two men left Columbus in a car on a long trip. One man drove at 44 miles per hour for 8 hours. The other man drove 32 miles an hour for 4 hours to complete the trip. What was the average speed for the trip?

Two men left Columbus in a car on a 500 mile trip. One man drove at 40 miles per hour for 8 hours. What was the average speed of the other driver if the trip was completed in 11 hours?

Two men left Columbus in a car on a long trip. One man drove at 50 miles per hour for 9 hours. The other man averaged 40 miles an hour for 6 hours to complete the trip. How far was the trip?
Two men left Columbus in a car on an 840 mile trip. One man drove at 60 miles per hour for 5 hours. What was the average speed of the other driver if the trip was completed in 17 hours?

Two men left Columbus in a car on a long trip. One man drove at 30 miles per hour for 10 hours. The other man drove 48 miles an hour for 5 hours to complete the trip. What was the average speed for the trip?

Two men left Columbus in a car on a long trip. One man drove at 42 miles per hour for 5 hours. The other man averaged 54 miles an hour for 15 hours to complete the trip. How far was the trip?
Two men left Columbus in a car on a long trip. One man drove at 60 miles per hour for 12 hours. The other man averaged 42 miles an hour for 6 hours to complete the trip. How far was the trip?

Two men left Columbus in a car on a long trip. One man drove at 40 miles per hour for 5 hours. The other man drove 60 miles an hour for 15 hours to complete the trip. What was the average speed for the trip?

Two men left Columbus in a car on a 235 mile trip. One man drove at 32 miles per hour for 5 hours. What was the average speed of the other driver if the trip was completed in 8 hours?
Two men left Columbus in a car on a long trip. One man drove at 48 miles per hour for 4 hours. The other man averaged 36 miles an hour for 8 hours to complete the trip. How far was the trip?

Two men left Columbus in a car on a 485 mile trip. One man drove at 45 miles per hour for 3 hours. What was the average speed of the other driver if the trip was completed in 10 hours?

Two men left Columbus in a car on a long trip. One man drove at 50 miles per hour for 9 hours. The other man drove 40 miles an hour for 6 hours to complete the trip. What was the average speed for the trip?
Two men left Columbus in a car on a long trip. One man drove at 50 miles per hour for 15 hours. The other man drove 30 miles an hour for 10 hours to complete the trip. What was the average speed for the trip?

Two men left Columbus in a car on a long trip. One man drove at 35 miles per hour for 8 hours. The other man averaged 55 miles an hour for 12 hours to complete the trip. How far was the trip?

Two men left Columbus in a car on a 400 mile trip. One man drove at 30 miles per hour for 6 hours. What was the average speed of the other driver if the trip was completed in 10 hours?
There are 16 boxes on a loading dock, some are red and the rest are green. The total weight of the boxes is 270 pounds. If each of the 6 red boxes weighs 15 pounds, how much does a green box weigh?

There are a total of 6 red and 9 green boxes on a loading dock. Each of the red boxes weighs 35 pounds. Each green box weighs 20 pounds. What is the total weight for the boxes?

There are a total of 15 red and 10 green boxes on a loading dock. Each of the red boxes weighs 13 pounds. Each green box weighs 8 pounds. What is the average weight for a box on the loading dock?
There are a total of 16 red and 8 green boxes on a loading dock. Each of the red boxes weighs 25 pounds. Each green box weighs 40 pounds. What is the average weight for a box on the loading dock?

There are 20 boxes on a loading dock, some are red and the rest are green. The total weight of the boxes is 600 pounds. If each of the 12 red boxes weighs 20 pounds, how much does a green box weigh?

There are a total of 5 red and 12 green boxes on a loading dock. Each of the red boxes weighs 30 pounds. Each green box weighs 40 pounds. What is the total weight for the boxes?
There are 20 boxes on a loading dock, some are red and the rest are green. The total weight of the boxes is 300 pounds. If each of the 5 red boxes weighs 12 pounds, how much does a green box weigh?

There are a total of 7 red and 14 green boxes on a loading dock. Each of the red boxes weighs 15 pounds. Each green box weighs 6 pounds. What is the average weight for a box on the loading dock?

There are a total of 5 red and 6 green boxes on a loading dock. Each of the red boxes weighs 20 pounds. Each green box weighs 32 pounds. What is the total weight for the boxes?
There are a total of 10 red and 7 green boxes on a loading dock. Each of the red boxes weighs 22 pounds. Each green box weighs 15 pounds. What is the total weight for the boxes?

There are a total of 12 red and 8 green boxes on a loading dock. Each of the red boxes weighs 10 pounds. Each green box weighs 20 pounds. What is the average weight for a box on the loading dock?

There are 25 boxes on a loading dock, some are red and the rest are green. The total weight of the boxes is 325 pounds. If each of the 15 red boxes weighs 11 pounds, how much does a green box weigh?
There are a total of 8 red and 11 green boxes on a loading dock. Each of the red boxes weighs 16 pounds. Each green box weighs 12 pounds. What is the total weight for the boxes?

There are 22 boxes on a loading dock, some are red and the rest are green. The total weight of the boxes is 164 pounds. If each of the 10 red boxes weighs 8 pounds, how much does a green box weigh?

There are a total of 6 red and 10 green boxes on a loading dock. Each of the red boxes weighs 7 pounds. Each green box weighs 15 pounds. What is the average weight for a box on the loading dock?
There are a total of 5 red and 10 green boxes on a loading dock. Each of the red boxes weighs 14 pounds. Each green box weighs 8 pounds. What is the average weight for a box on the loading dock?

There are a total of 6 red and 8 green boxes on a loading dock. Each of the red boxes weighs 25 pounds. Each green box weighs 30 pounds. What is the total weight for the boxes?

There are 23 boxes on a loading dock, some are red and the rest are green. The total weight of the boxes is 400 pounds. If each of the 11 red boxes weighs 20 pounds, how much does a green box weigh?
Two men left Columbus in a car on a 360 mile trip. One man drove at 44 miles per hour for 6 hours. What was the average speed of the other driver if the trip was completed in 9 hours?

Two men left Columbus in a car on a long trip. One man drove at 60 miles per hour for 15 hours. The other man averaged 45 miles an hour for 15 hours to complete the trip. How far was the trip?

Two men left Columbus in a car on a long trip. One man drove at 32 miles per hour for 4 hours. The other man drove 52 miles an hour for 6 hours to complete the trip. What was the average speed for the trip?
Two men left Columbus in a car on a long trip. One man drove at 44 miles per hour for 8 hours. The other man drove 32 miles an hour for 4 hours to complete the trip. What was the average speed for the trip?

Two men left Columbus in a car on a 500 mile trip. One man drove at 40 miles per hour for 8 hours. What was the average speed of the other driver if the trip was completed in 11 hours?

Two men left Columbus in a car on a long trip. One man drove at 50 miles per hour for 9 hours. The other man averaged 40 miles an hour for 6 hours to complete the trip. How far was the trip?
There are 20 boxes on a loading dock, some are red and the rest are green. The total weight of the boxes is 300 pounds. If each of the 5 red boxes weighs 12 pounds, how much does a green box weigh?

There are a total of 7 red and 14 green boxes on a loading dock. Each of the red boxes weighs 15 pounds. Each green box weighs 6 pounds. What is the average weight for a box on the loading dock?

There are a total of 5 red and 6 green boxes on a loading dock. Each of the red boxes weighs 20 pounds. Each green box weighs 32 pounds. What is the total weight for the boxes?
There are a total of 10 red and 7 green boxes on a loading dock. Each of the red boxes weighs 22 pounds. Each green box weighs 15 pounds. What is the total weight for the boxes?

There are a total of 12 red and 8 green boxes on a loading dock. Each of the red boxes weighs 10 pounds. Each green box weighs 20 pounds. What is the average weight for a box on the loading dock?

There are 25 boxes on a loading dock, some are red and the rest are green. The total weight of the boxes is 325 pounds. If each of the 15 red boxes weighs 11 pounds, how much does a green box weigh?
A bowler was in a weekend tournament. He averaged 210 for each game he bowled on Saturday and 240 for each game on Sunday. He bowled 8 games on Sunday and 12 games on Saturday. What was his total score for the tournament?

A bowler was in a weekend tournament. He had a 200 average for the 4 games he bowled on Saturday. He finished his bowling on Sunday and ended with a total of 2820 for the 12 game tournament. What was his average on Sunday?

A bowler was in a weekend tournament. He averaged 210 for each game he bowled on Saturday and 240 for each game on Sunday. He bowled 9 games on Sunday and 6 games on Saturday. What was his average score for the tournament?
A bowler was in a weekend tournament. He averaged 225 for each game he bowled on Saturday and 205 for each game on Sunday. He bowled 8 games on Sunday and 12 games on Saturday. What was his average score for the tournament?

A bowler was in a weekend tournament. He averaged 205 for each game he bowled on Saturday and 230 for each game on Sunday. He bowled 10 games on Sunday and 5 games on Saturday. What was his total score for the tournament?

A bowler was in a weekend tournament. He had a 230 average for the 12 games he bowled on Saturday. He finished his bowling on Sunday and ended with a total of 4680 for the 22 game tournament. What was his average on Sunday?
APPENDIX D

TEST ITEMS FOR EXPERIMENT II

A copy of the test items is provided on the next eight pages. The first four pages were given during the first testing session. The next four pages were given during the second testing session. The order of the pages was varied so that only one-fourth of the subjects received a test with the pages in this order.
Two men left Columbus in a car on a long trip. One man drove at 52 miles per hour for 5 hours. The other man drove 30 miles an hour for 6 hours to complete the trip. What was the average speed for the trip?

Two men left Columbus in a car on a 480 mile trip. One man drove at 35 miles per hour for 8 hours. What was the average speed of the other driver if the trip was completed in 12 hours?

Two men left Columbus in a car on a long trip. One man drove at 30 miles per hour for 5 hours. The other man averaged 60 miles an hour for 10 hours to complete the trip. How far was the trip?
There are a total of 8 red and 4 green boxes on a loading dock. Each of the red boxes weighs 20 pounds. Each green box weighs 35 pounds. What is the average weight for a box on the loading dock?

There are a total of 6 red and 4 green boxes on a loading dock. Each of the red boxes weighs 15 pounds. Each green box weighs 35 pounds. What is the total weight for the boxes?

There are 13 boxes on a loading dock, some are red and the rest are green. The total weight of the boxes is 240 pounds. If each of the 9 red boxes weighs 20 pounds, how much does a green box weigh?
Two men left Columbus in a car on a 502 mile trip. One man drove at 50 miles per hour for 5 hours. What was the average speed of the other driver if the trip was completed in 11 hours.

Two men left Columbus in a car on a long trip. One man drove at 30 miles per hour for 5 hours. The other man drove 54 miles an hour for 7 hours to complete the trip. What was the average speed for the trip?

Two men left Columbus in a car on a long trip. One man drove at 30 miles per hour for 11 hours. The other man averaged 55 miles an hour for 6 hours to complete the trip. How far was the trip?
There are 12 boxes on a loading dock, some are red and the rest are green. The total weight of the boxes is 300 pounds. If each of the 4 red boxes weighs 15 pounds, how much does a green box weigh?

There are a total of 6 red and 14 green boxes on a loading dock. Each of the red boxes weighs 20 pounds. Each green box weighs 30 pounds. What is the total weight for the boxes?

There are a total of 3 red and 9 green boxes on a loading dock. Each of the red boxes weighs 40 pounds. Each green box weighs 24 pounds. What is the average weight for a box on the loading dock?
A bowler was in a weekend tournament. He averaged 235 for each game he bowled on Saturday and 210 for each game on Sunday. He bowled 8 games on Sunday and 12 games on Saturday. What was his total score for the tournament?

A bowler was in a weekend tournament. He had a 200 average for the 12 games he bowled on Saturday. He finished his bowling on Sunday and ended with a total of 4320 for the 20 game tournament. What was his average on Sunday?

A bowler was in a weekend tournament. He averaged 215 for each game he bowled on Saturday and 230 for each game on Sunday. He bowled 4 games on Sunday and 8 games on Saturday. What was his average score for the tournament?
A man accepts a job scraping and painting a house. He is paid $18 a day for the 6 days that he scrapes the house. To both scrape and paint the house takes 14 days. He is paid $340 for this job. How much is he paid for each day that he paints the house?

A man accepts a job scraping and painting a house. He is paid $20 each day for scraping the house and $35 each day that he is painting. He works 12 days scraping and paints for 8 days. How much will he have earned on this job?

A man accepts a job scraping and painting a house. He scrapes the house for 4 days and is paid $15 a day. His pay is then changed to $24 per day for the 8 days that he paints the house. What is the average daily pay for this job?
A bowler was in a weekend tournament. He averaged 220 for each game he bowled on Saturday and 205 for each game on Sunday. He bowled 12 games on Sunday and 8 games on Saturday. What was his average score for the tournament?

A bowler was in a weekend tournament. He averaged 200 for each game he bowled on Saturday and 230 for each game on Sunday. He bowled 5 games on Sunday and 10 games on Saturday. What was his total score for the tournament?

A bowler was in a weekend tournament. He had a 210 average for the 5 games he bowled on Saturday. He finished his bowling on Sunday and ended with a total of 2340 for the 11 game tournament. What was his average on Sunday?
A man accepts a job scraping and painting a house. He scrapes the house for 6 days and is paid $15 a day. His pay is then changed to $25 per day for the 4 days that he paints the house. What is the average daily pay for this job?

A man accepts a job scraping and painting a house. He is paid $22 a day for the 10 days that he scrapes the house. To both scrape and paint the house takes 16 days. He is paid $400 for this job. How much is he paid for each day that he paints the house?

A man accepts a job scraping and painting a house. He is paid $15 each day for scraping the house and $35 each day that he is painting. He works 6 days scraping and paints for 9 days. How much will he have earned on this job?
APPENDIX E

SAMPLE INTERACTION BETWEEN RESEARCHER AND SUBJECTS
Each student received a handout with the sample problems listed. To clarify the small group treatments, a typical explanation has been presented below for a specific sample problem.

Sample Problem:

Two men left Columbus in a car on a long trip. One man drove at 40 miles per hour for 8 hours. The other man drove at 50 miles per hour so that the trip was completed in 14 hours. How far was the trip?

R. We are given a problem. Please read the problem to yourselves, while I read it aloud. "Two men left Columbus in a car on a long trip. One man drove at 40 miles per hour for 8 hours. The other man drove at 50 miles per hour so that the trip was completed in 14 hours. How far was the trip?" What are we supposed to determine in this problem?

S. How far the trip is?

R. Right! - We are given information about the speed and time that each man drove. How fast did the first driver travel?

S. 40 miles per hour.

R. How far did the first driver travel?

S. I don't know.

R. How many hours did the first driver travel.

S. 8
R. Right. Now we know that the first driver traveled 40 mph for 8 hours. How do we figure the distance the driver traveled?

S. Multiply.

R. Multiply which numbers?

S. 40 and 8

R. So we find the first man traveled 40 x 8 or 320 miles.

Pause (allow students to multiply 40 and 8)

R. We must still find the distance traveled by the second driver. Can we find the speed and length of time that the second man drove on this trip?

S. The speed is 50 miles per hour.

R. Right, but do we know the length of time?

S. 14 hours.

R. Wait, I think the trip is completed in a total of 14 hours. But this is the total amount of time needed to complete the trip. How many hours did the first man drive?

S. The first man drove 8 hours.

R. How long did the second man drive if the total time is 14 hours?

S. 6 hours.

R. So we have the second man driving 6 hours at 50 miles per hour. Therefore, the second man drove 50 x 6 or 300 miles. Do you remember how far the first man drove?
S. 320 miles.

R. So the first man drove for 320 miles and the second man drove 300 miles. How do we find the total milage?

S. Add.

R. Right we add 320 and 300 and get 620 miles.


166
Burns, P. C., & Yonally, J. L. Does the order of presentation of numerical data in multi-step arithmetic problems affect their difficulty? School Science and Mathematics, 1964, 64, 267-270.


Huffman, L. R., Burke, R. J., & Maier, N. R. Does training with differential reinforcement on similar problems help in solving a new problem? Psychological Reports, 1963, 13, 147-153.


Tennyson, R. D., Wooley, F. R., & Merrill, M. D. Examplar and non-examplar variables which produce correct classification behavior and specified classification errors. *Journal of Educational Psychology*, 1972b, 63, 144-152.


