

A STUDY TO DETERMINE THE
EFFECTIVENESS OF A TECHNIQUE
EMPLOYING AN AMBIGUOUS STIMULUS
FOR ASSESSING A CHILD'S LEVEL OF
SKILL AND CONCEPT DEVELOPMENT
IN THE AREAS OF ADDITION
AND SUBTRACTION

Dissertation for the Degree of Ph. D.
MICHIGAN STATE UNIVERSITY
JACQUELINE RESH LONG
1975



This is to certify that the

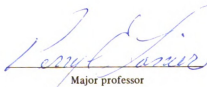
thesis entitled

A Study to Determine the Effectiveness
of a Technique Employing an Ambiguous Stimulus
for Assessing A Child's Level of Skill and
Concept Development in the Areas of Addition
and subtraction presented by

Jacqueline Resh Long

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Elementary Education


Major professor

Date 11/14/75



46

~~_____~~

208

11 ~~_____~~ 124

C495557

ABSTRACT

A STUDY TO DETERMINE THE EFFECTIVENESS OF A TECHNIQUE
EMPLOYING AN AMBIGUOUS STIMULUS FOR ASSESSING A
CHILD'S LEVEL OF SKILL AND CONCEPT DEVELOPMENT
IN THE AREAS OF ADDITION AND SUBTRACTION

By

Jacqueline Resh Long

The contributions of Skinner, Bruner, and Piaget have influenced new goals in education and new approaches to instruction. These new goals and approaches to instruction have created problems and needs for teachers.

A technique of evaluation was developed in pilot studies to help resolve the following problems and needs experienced by teachers in evaluating student learning:

1. Validate a method of measuring student achievement at the symbolic level of concept representation which would then open the way for researching this technique at the concrete and pictorial-diagrammatic levels of concept representation.
2. Drastically reduce the time required for preparing, adminis-
tering, and correcting tests.
3. Drastically reduce the time students would spend in being evaluated.
4. Offer a record of individualized growth by affording a teacher a collection of evaluations individually submitted which shows what a child regards as "hard" on a daily basis. This, then,

can be placed in a folder for the child, parent, or teacher to examine.

5. Place an emphasis on a child's ability to assess his own knowledge and recognize self-growth by asking him to submit an example of what he can do. This technique of evaluation is consistent with the goals of a behavioral philosophy of self and environmental assessment.

The purpose of this research is to evaluate the researched technique for assessing a child's level of skill and concept development in the areas of addition and subtraction. The assessment technique to be employed in this instance is limited to the symbolic representation of the mathematic's concepts and skills being examined. The limitation was placed on the study, because of the lack of instruments available in the concrete or pictorial-diagrammatic modes of concept representation with which to compare the newly researched technique. Currently accepted instruments of evaluation are tests primarily written to measure symbolic representation.

Several examiners used the technique in this study and administered the diagnostic tests to groups and individual children attending public schools. The testing technique employed an ambiguous verbal stimulus to which a child was asked to respond. The response of the student being evaluated was then correlated with a traditional diagnostic test written for this study for validation of the results. Using a Pearson product-moment correlation, a value of $r = .85$ for addition and $r = .81$ for subtraction was found. Constructing confidence

intervals for these two correlations ($P = .99$) ρ will be between .75 and .91 for addition and .66 and .90 for subtraction.

The following hypotheses were tested using a series of t-tests with an α level of .05 to determine if there were differences between groups in their ability to use the testing technique in this study.

1. There will be no significant differences between the high, average, and low achievers as determined by the Iowa Achievement tests in their ability to assess their level of abstract achievement.
2. There will be no significant differences between the high, average, and low achievers as determined by teacher judgment in their ability to assess their level of abstract achievement.
3. There will be no significant differences between Blacks and Caucasians in their ability to assess their level of abstract achievement.
4. There will be no significant differences between girls and boys in their ability to assess their level of abstract achievement.
5. There will be no significant differences between children from high, average, and low family incomes in their ability to assess their level of abstract achievement.

No significant differences between groups were noted. Therefore, it appears that all groups in the study can use the testing technique equally well.

The following hypothesis was tested to determine if there was a racial bias with respect to what a child perceives as "hard."

There will be no significant differences between racial groups in what they perceive as "hard."

A series of chi-square tests were used with an α level of .05. Holding achievement constant no racial bias was found with respect to what is considered "hard."

A STUDY TO DETERMINE THE EFFECTIVENESS OF A TECHNIQUE
EMPLOYING AN AMBIGUOUS STIMULUS FOR ASSESSING A
CHILD'S LEVEL OF SKILL AND CONCEPT DEVELOPMENT
IN THE AREAS OF ADDITION AND SUBTRACTION

By

Jacqueline Resh Long

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

College of Education

1975

TABLE OF CONTENTS

	Page
LIST OF TABLES	v
LIST OF FIGURES	vi
 Chapter	
I. INTRODUCTION	1
The Problem	2
The New Perspectives	2
The Effect on Curriculum	4
The Effect on Instruction	5
The Effect on Mathematics Instruction	7
The Effect on Teacher Roles	8
Resulting Problems and Needs for Teachers	9
Purpose of the Study	10
General Evaluation Procedures of the Partial Solution	12
Anticipated Outcomes of the Study	12
Assumptions	14
Limitations of the Study	14
Definition of Terms	15
The Pilot Studies	17
II. REVIEW OF THE LITERATURE	19
Introduction	19
Role of Evaluation in Teaching Models	20
Role of Evaluation in Mathematics	21
Historical Development of Standardized Testing	24
Historically Developed Criteria for Judging Evaluation Instruments and Measurements	33
Validity	34
Reliability	36
Usability	37
Review of the Research in Math Instruction	37
Evaluation Methods in Assessing Learning in a Math Lab	43
Anecdotal Records	43
Rating Scales	44
Checklists	45
Interview	46
Thresholding	47
The Use of Ambiguous Stimuli in Testing	50

Chapter	Page
III. PROCEDURE AND METHODOLOGY	53
Setting and Sample	53
Examiners	54
Instruments and Methods Used for Validating the Technique in This Study	55
Procedure	55
Methods of Analyzing Data	58
IV. PRESENTATION AND ANALYSIS OF THE DATA	65
Correlation Between the Technique in This Study and the Test Written for This Study	65
Child's Ability to Assess Himself	67
Analysis of the Data Concerning Hypotheses B1 through B5 of the Study	70
Analysis of the Data Concerning Hypothesis C in the Study	76
V. SUMMARY, GENERALIZATIONS, AND IMPLICATIONS FOR FUTURE RESEARCH	79
Criteria for Judging Testing Instruments and Measurements	79
Accuracy of a Child's Self-Assessment	86
Different Groups' Ability to Use the Testing Technique in This Study	89
A Racial Bias with Respect to What is "Hard"	89
Analysis of the Distribution of Percentage of Correct Response Scores with Respect to the Technique in This Study	89
A Review of the Stated Purpose of This Study	90
Implications for Future Research	92
Usability of the Technique in This Study in Other Areas	92
Areas of Mathematics Education to Be Researched Using the Technique in This Study	93
Appendix	
A. QUESTIONS USED IN PILOTS	96
B. PROCEDURE HANDOUT	97
C. TESTS	99
BIBLIOGRAPHY	106

LIST OF TABLES

Table	Page
1. Summary of the effect of activity and model methodologies on the learning of mathematics in kindergarten through third grade	39
2. Summary of studies to determine the effectiveness of teaching with models and activities in grades four through six	40
3. Summary of the effect of activity and model methodologies on the learning of mathematics in grades seven through twelve	42
4. F-tests for determining the differences in variance of the groups in the study	72
5. Group differences in their ability to use the testing technique in this study	74
6. Summary of the results of the chi-square tests with addition	77
7. Summary of the results of the chi-square tests with subtraction	78

LIST OF FIGURES

Figure	Page
1. Scattergram of the results of the test written for this study and the technique of this study	60
2. Spread of scores for addition and subtraction	85

CHAPTER I

INTRODUCTION

Recent acceptance of theories in the science of behavior, cognitive development, and concept representations have created new approaches to instruction. These, in turn, have created new problems and needs for teachers. To better understand the dimensions of the situation, this chapter will cover the following topics:

1. the new perspectives and their corresponding effect on curriculum, instruction, and teacher roles, and the resultant problems and needs that have arisen for teachers of mathematics;
2. a description of the purpose of this study, which attempts to identify a partial solution to one of the problems.
3. a description of the general procedures that were undertaken to evaluate the partial solution, including the procedure for both administering and evaluating the technique;
4. a discussion of the anticipated outcomes of the study;
5. a presentation of the assumptions that undergird the research, the limitations of the research, and the definitions of key terms employed in this study; and
6. an examination of the pilot studies which helped to develop the technique.

The Problem

The New Perspectives

The new perspectives affecting educational goals have their origin in the recently defined nature of man. The simplistic view theorized by B. F. Skinner offers man an opportunity to achieve a relative freedom heretofore unknown to him because of his past ignorance and refusal to recognize the factors in his environment which limit or destroy his freedom. Skinner, contrary to the generally accepted theory of internal control, has hypothesized that man is born with a differentiated ability to respond to stimuli, and through continuous conditioning the probability for any given behavior is changed. Acceptance of the concept that behavior arises primarily from conditioning requires that man learn to assess which environmental factors affect him, and in what way, before he can achieve maximum freedom from environmental control.

Skinner has also contributed a method of determining relationships between man and his environment through the observation of behavior, its stimuli, and reinforcers, without theorizing about unobservable factors. Thus, any individual with skill in assessing his milieu is able to determine the behavioral cause-and-effect relationships that exist for him, personally, and thereby possibly change the portions of his environment which adversely affect his desired behavior.

The work of Jerome Bruner, Jean Piaget, and many math educators has clearly demonstrated that learning needs to begin with concrete models and progress to symbolic models. Van Engen (1949), supporting

the theories of both Bruner and Piaget, pointed out that the "meaning of words cannot be thrown back on the meaning of other words. When the child has seen the action and performed the act for himself, he is ready for the symbol for the act."

Piaget has been the major contributor of theoretical support for the use of concrete before symbolic models. He has proposed a comprehensive theory of cognitive development that encompasses individual growth from birth to maturity. Fennema (1972) describes Piaget's concept:

According to Piaget's theory, schemas (mental structures) are formed by a continual process of accommodation to and assimilation of the individual's environment. This adaptation (accommodation and assimilation) is possible because of the actions performed by the individual upon his environment. These actions change in character and progress from overt, sensory actions done almost completely outside the individual to partially internalized actions that can be done with symbols representing previous actions, to completely abstract thought done entirely with symbols. This development in cognitive growth involves, first the use of physical actions to form schemas. Learners change from a predominant reliance on physical action to a predominant reliance on symbols.

Bruner has theorized that a learner utilizes, in order, three representations in the process of acquiring a given concept. The first is the enactive or manipulative stage in which an understanding of a concept can be gained only as far as the actions in correspondence to an object possess the attribute of the idea to be learned. In the second stage, ikonic representation, a child can represent the world by an image of the original object or action performed on the object, without the object being present. The final representation is symbolic.

The Effect on Curriculum

Educating an individual both formally and informally to live effectively within society has been the primary role of schools. Unfortunately, past efforts have entailed the imparting of "factual" knowledge without emphasizing the origin of these facts, thereby concealing the structure of the subject area studied. Hilda Taba (1967) is critical of a curriculum emphasizing the learning of facts without structuring their implications: "Because specific facts become obsolete more rapidly than basic concepts or main ideas, they are not significant in themselves. Their chief function is to explain, illustrate, and develop main ideas."

Bruner (1960), by pointing out the historic problem of how to teach the basic structure of a subject area, gives evidence of the cafeteria style, fact-teaching of the past. He maintains that since so little is known about teaching the fundamental structure, facts rather than structure have been emphasized in the education of an individual.

Studies done by Lankford (1974), Swart (1974), and Peck and Jencks (1974) have attempted to determine what is being taught in today's traditional math classes. These studies found classrooms of children memorizing number facts, definitions, rules, and algorithms.

A curriculum consistent with a behavioral oriented philosophy of education that is behavior oriented should have an emphasis which fosters its goals. The education of an individual should now afford him the opportunity to develop the skills necessary to maximize his

ability to perceive cause-and-effect relationships by helping him to order and structure his milieu, thus enabling him to become as independent as possible of both his physical and human environments. The essence of this freedom remains less than absolute because of man's inability to exist outside of an environment with controlling stimuli and reinforcers. John Holt, in Freedom and Beyond, refers to man's relative freedom as a constrained life.

We are all and always constrained, bound in, limited by a great many things, not least of all the fact that we are mortal. We are limited by our animal nature, by our model of reality, by our relations with other people, by our hopes and fears.

This "constrained" life can only have an individually achieved maximum freedom based on an individual's unique genetic make-up and unique sequence of experiences.

The Effect on Instruction

Fennema (1972), in summarizing a multitude of studies which tended to support Piaget's theory of cognitive development and Bruner's theory of concept representation, states:

Collectively, these data tend to support the hypothesis that a learning environment embodying representational models suited to the developmental level of the learner facilitates learning better than a learning environment that ignores the developmental level of the learner.

The acceptance of Bruner and Piaget's theories suggests that models be present in a learning environment if conceptual learning is to take place. Through the use of such models each child would be able to test the correctness of his perceived generalizations for himself or with other students, thereby placing the authority for learning on each

child or his group. This type of learning environment would foster individual growth in the ability to perceive relationships and encourage a child to be dependent on his own perceptions rather than on those of a teacher or some other authority. The child is thus weaned from his dependent state to one of independence.

Taba (1967) states:

In order to develop autonomy of thought, students need opportunities to organize their own conceptual systems and to develop their skills for independent processing of information. Consequently, the nature and the organization of learning experiences should be calculated to encourage the learner to inquire, to do his own thinking, to develop his own ways of working out problems, and to try out his own ideas. Faced with the temptation to provide the answers and solutions, the teacher must grant the learner the right to come to grips with the learning process, even though the products may be less refined than the teacher would wish.

Skinner's postulation that individuals are born with a differentiated ability to respond to stimuli, Piaget's theory of cognitive development, and Bruner's theorized stages of concept representation all point out a need for the individualization of instruction. By postulating a genetic component to individual response, a uniqueness of response is implied. Piaget's theorized stages of cognitive development and Bruner's modes of representation also imply a variety of levels of cognitive functioning and modes of concept representation within any given group of children, necessitating the creation of a learning environment which offers a variety of learning situations designed to accommodate the uniqueness of each individual.

This individualized instruction could be achieved within a classroom laboratory with concrete, pictorial, or diagrammatic and

symbolic models for the children to use in the attainment of concepts. Each child would use a model most meaningful to him and would progress at his own pace. The concepts to be learned could be determined for the child by his teacher with a sequenced exposure to models to ensure the eventual learning of the concept, or a nondirected laboratory exposure to large collections of models could be used. Students in this type of a milieu can grow in their ability to learn through student interactions which could broaden their perceptions, or they can learn through solitary experimenting. Both of these situations permit individuals to differ in the selection of meaningful models and in their ability to perceive relationships while being a member of a learning group. Lab-oriented experiences which use individual or small group explorations, with materials and teachers as resources, would foster the type of learning situation consistent with the goal of teaching children how to perceive relationships.

The Effect on Mathematics Instruction

The unique contribution which mathematics instruction offers to the education of a person is the opportunity to observe relationships directly through the use of mathematical models which range from the concrete to the symbolic. A concrete model (Fennema, 1972) represents a mathematical idea by means of three-dimensional objects. A second type of model is the pictorial or diagrammatic. Through pictures or diagrams, the attributes of certain mathematical concepts are demonstrated. Finally, symbolic models represent a mathematical idea by means of commonly accepted numerals and signs that denote mathematical

operations or relationships. From the use of such models children and adults can experience the act of learning to learn in a math laboratory with models which encourage growth in skills of observing, systematizing, formulating, and testing generalizations. Mathematics also offers the opportunity to develop the ability to quantify data and tersely express relationships symbolically, so that patterns in any given situation can be discerned more easily. These skills are very necessary if individuals are to develop to their fullest capacity their competency to determine cause-and-effect relationships.

The Effect on Teacher Roles

The role of the teacher in instruction can contribute to or hinder the achievement of the educational goal of independence, for the product or consequence of this instruction is a function of this role and can be freeing or restricting with respect to an individual's growth.

In the traditional instructional milieu, where authority for learning rests solely with the instructor, two interrelated conditions arise. First, a student becomes dependent on his instructor for the "rightness" or "wrongness" of his generalizations rather than on his own ability to prove to himself the truth of his conclusions. Second, a student is limited by his instructor's knowledge rather than his own concerning the relationships it is possible for him to perceive, and he is then limited to perceiving only those relationships which his teacher relates to him. Therefore, traditional expository teaching violates the goal for achieving a maximum amount of independence for

each individual by limiting learning and forcing an individual to depend on the perception of others. For similar reasons, programmed instruction in areas of concept development and guided discovery where only one outcome is acceptable are also deterrents to the goal of independence.

Resulting Problems and Needs for Teachers

The contributions of Skinner, Bruner, and Piaget have influenced new goals in education and new approaches to instruction. Some of the problems and needs which have resulted from these changes are the following:

1. Teachers will be using a method of teaching that was not used with them.
2. Teachers will need to learn how and when to use models in their instruction.
3. Teachers will need to determine the student's stage of development, as defined by Piaget, and the appropriate model for depicting a particular concept best suited to the intellectual needs of the student.
4. Teachers will need to find models for concepts that they wish to teach and all the modes of representation for these concepts.
5. Teachers will need to learn how to organize their teaching days so that they can offer individualized instruction.
6. Teachers will need a system of daily record keeping to enable individual growth to be discerned and planned for.

7. Inherent in any teaching situation, especially an individualized lab approach to teaching mathematics, is the problem of accurately assessing the entering skill and mode of concept representation for each student. In addition, an accurate evaluation following each learning experience to redetermine the functioning level of the student must be made.
8. Teachers will need sizable amounts of time to prepare, administer, and grade tests for the myriad of levels in an individualized lab milieu. Instructional time will be significantly affected.
9. Teachers will need to set aside sizable amounts of student time for taking tests.
10. Teachers will have to find commercial tests or design their own to measure the concrete and pictorial-diagrammatic levels of concept representation. Presently, most accepted evaluation instruments test primarily the symbolic level of concept representation.

Purpose of the Study

A technique of evaluation was developed in pilot studies by this investigator which intended to do the following:

1. Validate a method of measuring student achievement at the symbolic level of concept representation which would then open the way for researching this technique at the concrete and pictorial-diagrammatic levels of concept representation.

2. Drastically reduce the time required for preparing, administering, and correcting tests.
3. Drastically reduce the time students would spend in being evaluated.
4. Offer a record of individualized growth by affording a teacher a collection of evaluations individually submitted which shows what work a child regards as difficult on a day-to-day basis. This, then, can be placed in a folder for the child, parent, or teacher to examine.
5. Place an emphasis on a child's ability to assess his own knowledge and recognize self-growth by asking him to submit an example of what he can do. This technique of evaluation is consistent with the goals of a behavioral philosophy of self--and environmental assessment.

The purpose of this research is to evaluate the pilot technique for assessing a child's level of skill and concept development in addition and subtraction. The assessment technique to be employed in this research is limited to the symbolic representation of the mathematics concepts and skills being examined. This limitation was placed on the study because of the lack of instruments available in the concrete or pictorial-diagrammatic modes of concept representation with which to compare the results of the technique in this study.

The lack of such instruments was established by requesting and subsequently reviewing the commercial diagnostic and achievement tests cited in the twenty-sixth yearbook *Evaluation in Mathematics*, of the

NC

br

Ma

al

or

re

sc

in

sy

ad

in

ch

co

st

wn

a

NCTM (National Council of Teachers of Mathematics) and the NCTM brochure, "Mathematics Tests Available in the United States." Marily Suydam's annotated list of unpublished evaluation instruments also was reviewed.

Since the concrete stage of concept representation is entirely omitted from all test items, and since the pictorial-diagrammatic representation is omitted from all tests for middle and upper elementary schools for most concepts, it is apparent that currently accepted instruments of evaluation are tests primarily written to measure the symbolic representation of concepts and skills.

General Evaluation Procedures of the Partial Solution

Pre- and in-service teachers used the pilot technique and administered the diagnostic test written for the study to groups and individual children attending public schools.

The test employed an ambiguous verbal stimulus to which a child was asked to respond. This response was evaluated and then correlated with the index of the diagnostic test written for this study to validate the results.

Anticipated Outcomes of the Study

The following major hypothesis will be tested to determine whether or not there is a correlation between the results of testing a child by a diagnostic test and the testing technique being studied:

- A There will be a high correlation between the results of testing using a diagnostic test and the results of testing using the technique being studied.

The following five hypotheses will be tested to determine whether or not there is a difference between groups in their ability to use the testing technique in this study.

- B1 There will be no significant differences between the high, average, and low achievers as determined by the Iowa Achievement tests in their ability to assess their level of abstract achievement.
- B2 There will be no significant differences between the high, average, and low achievers as determined by teacher judgment in their ability to assess their level of abstract achievement.
- B3 There will be no significant differences between Blacks and Caucasians in their ability to assess their level of abstract achievement.
- B4 There will be no significant differences between girls and boys in their ability to assess their level of abstract achievement.
- B5 There will be no significant differences between children from high, average, and low income families in their ability to assess their level of abstract achievement.

The following hypothesis will be tested to determine whether or not there is a racial bias with respect to what a child perceives as difficult or "hard."

- C. There will be no significant differences between racial groups in what they perceive as "hard."

Assumptions

Evaluation in mathematics instruction is based on several assumptions. First, determination of a student's stage of cognitive and mathematical development is a necessary task, regardless of the teaching model being used. Second, current diagnostic tests are relatively accurate in determining a student's competency level with abstract models of concept representation. Third, thresholding is a valid means of determining a level of students' functioning when using diagnostic tests. Fourth, proper sequencing of levels within a diagnostic test is necessary if thresholding is to be used as a means of determining a level of functioning. Fifth, there are three stages of concept representation: the concrete, pictorial-diagrammatic, and abstract.

Limitations of the Study

Three major limitations of this study should be noted. First, only two of the four operations with whole numbers were used in the study, and, no other areas of mathematics which might be assessed by the technique being evaluated will be researched. Second, the abstract stage of concept representation is the only stage considered because of

the problem of validating testing results for the concrete and pictorial-diagrammatic stages. Finally, only children in grades 1 through 6 were studied.

Definition of Terms

In what follows, the major terms used in this study are defined.

abstract (symbolic) models: Models which represent a mathematical idea by means of commonly accepted numerals and signs that denote mathematical operations or relationships.

ambiguous stimulus: A stimulus which elicits a variable response from a group of individuals.

behaviorism: The science of behavior which is attempting to understand the relationships between and within the genetic endowment, historical environment, and present environment of individuals with the ultimate goal of accuracy in the prediction of behavior.

commercial tests: Those tests prepared by various companies which attempt to measure mathematics achievement.

concrete models: Models which represent a mathematical idea by means of three-dimensional objects.

level of concept development: The level of model needed by a person in order to attain the concept being presented. The model representations are the concrete, pictorial-diagrammatic, and symbolic.

math lab milieu: A math learning environment having models that represent mathematical ideas concretely, pictorially-diagrammatically, and symbolically and a variety of instructional media, such as tape recorders, to enhance the learning of mathematics in an individualized or small group situation.

pictorial-diagrammatic models: Models which represent a diagrammatic mathematical idea by means of pictures, diagrams, or devices such as a number line, which illustrates many of the attributes of the idea.

proper sequencing: A sequencing of response categories which consists of an "ascending" series carried far enough to locate the transition part or threshold from one response category to another.

quantitative understanding: The understanding that comes with numerals, mathematical symbols, and operations which enables a child to relate these mathematical ideas to his environment.

teacher prepared tests: Those tests prepared by a teacher to measure the entering or terminal behavior of a student in mathematics.

teaching model: A set of associated ideas and concepts more or less organized around a larger conception of what teaching should be like. It enumerates the components of a teaching situation and shows a general relationship between these components.

testing technique: A method of eliciting student responses which indicates an achievement level without utilizing an instrument or prepared list of objective questions.

the

Mi

El

at

re

sc

"h

ef

re

an

le

in

ra

ze

re

e

thresholding: A level (threshold) of functioning ascertained by observing where in a sequenced task a person begins to make more errors than correct responses, or where this individual stops participating in the task.

The Pilot Studies

Pilot studies were conducted in Cornell School in Okemos, Michigan, and in several schools in the Lansing, Michigan, area by Elementary Intern Program students. Additional data were collected at Ball State University by students in methods classes who are required to tutor individual or small groups of elementary students.

These studies attempted to find out whether or not elementary school children would respond to an open question posed in terms of "hardness." Several forms of questions were used to determine the most effective. See Appendix A for the questions used.

Many children in the pilot studies conducted for this research responded to the assessment questions by giving a memorized problem and answer, that is, $2000 + 2000 = 4000$, $100 + 100 = 200$. Since the problems always used large numbers, it would appear that this behavior was intended to impress the examiner. To overcome this problem in the validation study, youngsters were asked to write a problem without zeros. This change in procedure appeared to give more dependable results. Requesting a child to check his results with an aid also eliminated memorized responses.

In addition, the pilot studies showed that, with further testing, a child who would not submit a problem was not able to respond to symbolic representation in the area being assessed. However, if this nonrespondent was given a manipulative aid of his choice, he could provide both problems and solutions.

Across operations, children indicated that "hardness" was equivalent to large numbers. The majority (43 out of 72) gave examples of "hard" problems using numbers greater than 100. When children were given mathematical models to use, their responses seemed to be correlated to the device used. If a model was used which limited the size of numbers to a quantity under 70, then the hardest problems submitted included numbers close to 70. If, as in Chip Trading, problems with regrouping were treated no differently than those without, children rarely cited problems with regrouping as "hard." These observations were made with only 22 students.

Finally, the pilot studies revealed that errors in posing assessment questions and interpretation of questions by children resulted in some children offering problems that they could not solve. These problems were generally solvable by the child who submitted the problem after a short period of instruction.

CHAPTER II

REVIEW OF THE LITERATURE

Introduction

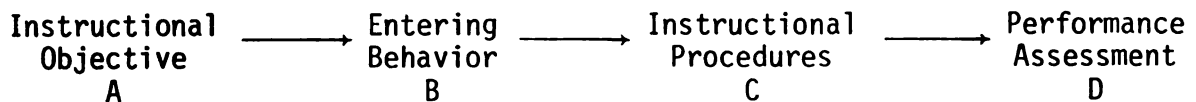
A review of the literature was made to establish the important role of evaluation within the theoretical framework of teaching models in general and within the teaching of mathematics in particular. In this chapter, a review of the development of standardized tests disclosing the historically based need for objective evaluation to ascertain a level of student cognitive functioning is followed by a presentation of the historically established criteria for judging evaluation instruments and measurements.

After examining the research conducted to determine the effectiveness of teaching mathematics using concrete, pictorial-diagrammatic, or symbolic models in a mathematics laboratory, numerous ways of evaluating learning in a mathematics laboratory which are currently being used are presented. The chapter concludes with a theoretical basis for employing a thresholding technique, followed by a presentation of the historical precedent of using an ambiguous stimulus in testing, as is employed in the testing technique in this study.

Role of Evaluation in Teaching Models

Both the behavior-modification teaching model and the discovery-learning model consist of a set of associated ideas and concepts more or less organized around a larger conception of what teaching should be and how it should be viewed. Nutshall and Snook (1973) have described the behavior-modification model: "[It] consists of that set of concepts and claims about teaching which has arisen from the attempt to apply the interpretive framework of behavioral psychology to the classroom." They add that "the discovery-learning model incorporates those views of teaching which place greatest emphasis on the self-directed activity of the student."

Glaser (1962) has developed a simple basic teaching model including the four essential components of any teaching situation. DeCecco (1968) pointed out that these components are present in most teaching models, especially in the models used to depict behavior-modification and discovery-learning. A basic teaching model (Glaser, 1962) is as follows:



Instructional objectives are measurable goals which a student should obtain by the completion of a segment of instruction. Entering behavior describes the student's level of cognitive and affective development prior to instruction. Instructional procedures refer to the input of a teacher in the changing of a student's behavior and is commonly called learning or achievement. Performance assessment consists of tests and observations used to determine how well the student has achieved the instructional objectives. Two of the four elements of the basic

model require that information from the student be collected. In noting the entering behavior all past experiences of a student deemed relevant to the new teaching situation must be assessed, while performance assessment in the portion of the model which deals with determining what learning took place with respect to the instructional objectives.

It is apparent from the literature that the evaluation of student learning is a necessary component of most teaching models. For the two models consistent with a behavioral philosophy, the behavioral-modification and the discovery-learning models, evaluation has a definite role.

Role of Evaluation in Mathematics

In the NCTM's twenty-sixth yearbook, Evaluation in Mathematics, Suelztz states emphatically the role of evaluation in mathematics:

Mathematics is an important part of the curriculum at all school levels beginning in the kindergarten. It is organized in a sequence of topics and activities that are associated with appropriate levels of maturity and ability of the students. Evaluation can identify and define steps and levels in the sequence that are appropriate for a given grade or age level. Careful evaluation should show not only how far a pupil has progressed in the major steps of a sequence, but also how well he has understood and mastered a particular step. Good evaluation will show the facts and skills mastered (and those not mastered) by the student, his attitude toward the subject, and the depth of understanding and insight accompanying his work.

He adds:

Evaluation is useful in determining the relative ease or difficulty of learning, applying, or remembering a topic, and materials. We need to know how long it takes to master a given concept, the suitable concepts for different grades, the appropriate sequence of concepts, and the aids the teacher needs to build mastery of each concept.

Reisman (1972) points out the importance of evaluation in determining a mathematics curriculum for effective instruction. By ascertaining a student's level of functioning, a curriculum can be developed which will meet the needs of the students involved without the negative ramification of an inappropriate curriculum. She states:

In looking at the mathematics curriculum, one must consider the level of difficulty involved. If the curriculum contains an abundance of material which is too advanced or too difficult for the student, he may become frustrated and give up; on the other hand, a curriculum that is too easy leads to boredom and the student again may give up.

Reys (1971), in an article on manipulative materials, remarked that to judge the effectiveness of materials, it would be wise to evaluate learning following their use.

Do evaluate the effectiveness of materials after using them. Immediately upon the completion of an activity, it can be very helpful to note particular problem areas, strengths, weaknesses, and suggestions and to define areas of needed improvement as well as possible areas of modification. A continuous reevaluation of manipulative materials ultimately results in better materials as well as more effective use of them.

Ewbank (1971), in an article on mathematics labs, discussed the inherent problem of evaluating mathematics learning in a laboratory milieu.

Some people use standard methods, that is, teacher-made or standardized tests. But the results of these tests may be deceptive, as it is very difficult to measure understanding and grasp of concepts in this way. . . . One way to measure progress in the mathematics laboratory is to look at the quality of written reports. . . . A high standard of written reports should be required, but in the primary grades it is a mistake to force children to write reports until they are ready to do so.

Mathematics can be learned by manipulating devices such as the equalizer balance and colored rods without any writing at all. Small children need to play with containers of sand, water, and so on, and in the process they grasp very important concepts such as the conservation of quantity. I do not see how you can evaluate this in the orthodox way. . . . For children at this early stage of development, subjective evaluation may be the best means. However, subjective evaluation should be based on the teacher's notes, anecdotal records, and a scrutiny of the child's progress in his written recording. Short periodic quizzes may be useful to show up those who cannot do certain processes or who obviously have not grasped the relevant concepts.

Sueltz summarizes the basic functions of evaluation in the total mathematics program in the following way:

1. Evaluation can establish levels of learning and locate a student at a level suitable for his current status in mathematics.
2. Evaluation is useful in improving the mathematics program in terms of curriculum, content, and organization, selection of materials for learning, and modes of instruction and learning. It can furnish data which should be used in making value judgments.
3. The place of mathematics in modern society can be studied and appraised in its many ramifications, and the results of such appraisal can be used in an appreciative way and also as a factor in determining the curriculum.
4. Competent evaluation of the mathematics program of a school is useful in keeping the clientele of the school informed and in answering questions raised by critics.
5. The information and data collected in evaluation form the substance of a student's record in school. These data are useful not only for records and reports, but also for research.
6. Evaluation is much concerned with helping the student learn mathematics more effectively. Hence, it seeks answers to many questions dealing with the kind of mathematics, the level of learning, motivation, and aspiration.
7. Different modes of learning and their effectiveness when applied to mathematics should be evaluated. This applies to various types of materials, various levels of learning, and various types of students.
8. Finally, evaluation itself provides valuable learning experiences that a good teacher will capitalize on to enhance the work of the students.

The importance of evaluation in mathematics education is clearly stated in Suelztz's summary. In order accurately to perform the evaluations he cites, new instruments and techniques of evaluation will have to be developed which take into account Piaget's theory of cognitive growth and Bruner's theory of concept representation within a behavioral philosophy of education.

Historical Development of Standardized Testing

A review of the historical development of testing reveals that testing did not originate in the pursuit of educational ideals, but, rather, stemmed from personal and political considerations. Mehrens and Lehman have described the historical setting:

When Binet developed his first scale, he was concerned with devising a means of removing dull pupils from the overcrowded schools in Paris rather than with constructing an instrument specifically designed to help the classroom teacher relate certain intellectual qualities to the learning process. Horace Mann really did not intend to devise an objective measure of pupil accomplishment. His criticism of the public schools in Massachusetts infuriated a group of teachers and lay citizens in Boston. This group were intent in resisting and refuting Mann's opinions. In the end, as a solution to the problem, it was agreed to prepare written examination questions in history, geography, vocabulary, science, arithmetic, astronomy, and grammar. This survey instigated by Horace Mann, was the first instance in which the same written examination was given to a sample of all pupils at the same school level, and where the papers were scored under uniform conditions. Although the findings confirmed Mann's contention that the public schools were not as good as claimed, it would appear that the findings did not serve as a stimulus to more objective and refined evaluation techniques in American public schools.

Green (1970) noted the following about Mann's achievement tests:

It is interesting to note that these same examinations were given to all eighth graders in the Boston schools following World War I in order to compare the results with the scores of the original pupils. The children in 1919 excelled their 1845 predecessors by a considerable degree in all areas except arithmetic problem solving. Another examination given in Springfield, Massachusetts, in 1846, and a retest in 1906 gave results similar to those in Boston (Lubberley, 1934).

At the time of the American Civil War a little known man in the field of education constructed the first objective educational test. Reverend George Fisher, an English schoolmaster, devised a series of tests to measure accomplishment in spelling, grammar, handwriting, composition, mathematics, and other school subjects. This series of tests was referred to as a Scale Book. Mehrens and Lehmann (1969) have described its contents.

Thorndike made a major contribution in 1904 when he published the first comprehensive book in the field, Mental and Social Measurement. In this book he proposed several of the principles which are still used in constructing standardized tests. Among these principles were (1) test items should be scaled according to difficulty, (2) tests should be objectively scored, and (3) tests should have statistical norms. Thorndike gave further impetus to the field by publishing the 1909 "Scale for Handwriting of Children" and by encouraging students to do further work in the field. During this period there were several new tests which helped turn the tide of schoolmen in favor of the movement. These tests included C. W. Stone's 1908 edition of a standardized achievement test in arithmetic, the arithmetic scales by Courtes in 1910, and the "Composition Scale" by Ayres in 1912.

The impetus for the continued development of standardized tests came from three sources: (1) unreliability of school marks as an indicator of school achievement, (2) a group of city school surveys conducted between 1910 and 1917 in which standardized tests were used to measure student achievement, and (3) the results of three noteworthy studies.

Mehrens and Lehmann (1969) have pointed out the problem of unreliable teacher grading:

In 1912 and 1913, Storch and Elliott had a group of teachers independently grade an English essay, a geometry paper, and a history paper. They found considerable variation in grades assigned (even with the geometry paper, which we would assume to be more amenable to objective evaluation). In 1928, Falls had 100 English teachers grade an essay written by a high school senior (who, incidently, wrote for a newspaper). The teachers were required to assign both a numerical grade to the essay as well as indicate the grade level of the student. Once again, as in Storch and Elliott's study, there was marked variation in both the numerical grades assigned and the estimated grade level of the writer. The grades varied from 60 to 98 percent and the grade level from 5 to 15. These kinds of studies led to the search for, and development of, more objective procedures for testing and grading students.

From the school surveys done using standardized tests, the economic value of producing an acceptable test battery became apparent. In 1919 the Stanford Achievement Battery was published. It was designed primarily for use at the elementary level. Green (1970) has stated:

Although achievement tests changed very little after the publication of this battery, numerous test publishing companies were established, and standardized tests were developed in all fields. An idea of the rapid expansion in the field can be gained from Hildreth's bibliography of mental tests and rating scales. Hildreth listed 3500 titles in 1935, 4279 titles in 1939, and 5294 titles in 1945.

Three influential studies which showed the major development in standardized achievement testing in the 1940s and 1950s as listed by Mehrens and Lehmann (1969), were: (1) the Eight-Year Study of the Progressive Education Association in 1942, (2) the College Entrance Examination Board long-range study initiated in 1952, and (3) the Cooperative Study of Evaluation in General Education completed in 1954.

These studies showed an increased use of standardized achievement tests in our public schools, a beginning inclusion of critical thinking, application of knowledge, synthesis, and evaluation, and the refinement of techniques used to construct and standardized achievement tests.

Ayres (1918) prophesied the importance and subsequent growth of the educational measurement movement in the seventeenth yearbook of the National Society for the Study of Education, Part II: "Knowledge is replacing opinion, and evidence is supporting guesswork in education as in every other field of human activity." In the final chapter of that yearbook, Judd (1918) noted:

The time is rapidly passing when the reformer can praise his new devices and offer as the reason for his satisfaction, his personal observation of what was accomplished. The superintendent who reports to his board on the basis of mere opinion is rapidly becoming a relic of an earlier unscientific age. There are indications that even the principals of elementary schools are beginning to study their schools by exact methods and are basing their supervision on the results of their measurements of what teachers accomplish.

Merwin (1969) pointed out that the changes in educational evaluation have evolved through interaction with (1) accepted theories and practices of education, (2) the role accepted for evaluation in the educational process, and (3) technical developments in educational evaluation.

Dobbin (1956), citing evidence of the effect of learning theories and practices in education on evaluation, noted that not only fundamental changes in learning theory, but also sweeping changes in enrollment and school organization patterns, have led to changing concepts of assessing achievement since the early 1930s. Starch (1916)

suggested that evaluation concern itself with determining individual differences in what pupils learn. Educational practices which evolved from this general idea ranged from "homogeneous grouping" to and including individualized instruction. Dressel (1950) pointed out that testing cannot avoid influencing instruction.

The role of evaluation in educational changes and the resultant changes in evaluation require examination.

1. The role in general school planning.--Efforts by Haggerty (1917) to determine the effect of evaluation on school planning gave evidence that, as a result of testing, changes occurred in (a) classification of pupils, (b) school organization, (c) courses of study, (d) methods of instruction, (e) time devoted to subject, and (4) methods of supervision. Twenty years later, Reaves commented: "The development of the measuring movement and the perfection of tests for the measurement of achievement and mental capacity have made possible great advances in educational administration."

2. The role in instruction.--Merwin (1969) pointed out that during the 1930s there were a number of proposals suggesting that school testing programs should be conducted in the fall of the year as a basis for evaluating the level of achievement following instruction. Troyer (1947) proposed that pretesting be used to determine the degree of knowledge and skills the students possessed which were prerequisites to the concepts to be taught. In the forty-fifth yearbook, Douglass and Spitzer wrote: "For many years we believed that good teaching begins where the child is, at the point to which his achievement has

brought him. We realize that we must take into consideration what the pupil already knows if we are to guide his learning from then on in an effective manner."

3. The role of student decision making.--Simpson (1953)

cogently argued that most learning takes place outside the classroom and that much more learning could take place if students developed skills for realistically planning and evaluating their own educational experiences.

4. Changing concepts and the content of evaluation.--

(a) Merwin (1969) pointed out that we apparently are in the process of completing a cycle approximately fifty years in length. Monroe's book, Measuring the Results of Teaching, described evaluation as focusing on very detailed objectives related to skills. Glaser (1967), at the Invitational Conference on Testing Problems, presented graphical descriptions of the accomplishments of individual students over time on relatively minute units of learning. Between the publications of these two reports, there has been considerable emphasis on more general outcomes. (b) Acceptance of the philosophical position that the teacher should take each child "where he is" and move him as far as possible toward his maximum potential development calls for a measure of status at two points in time as a basis for determining change, or "growth." (c) Bloom (1956) gave considerable impetus to the broadening of evaluation efforts to include the measuring of "higher mental processes." A publication by Krathwohl, Bloom, and Masia (1964) holds promise for broadening evaluation procedures to take into account very important educational objectives that fall in the

affective area. Environmental factors affecting learning have long been recognized, but only in recent years, with the work of Pace and Stern (1959), Wolf (1965), and Coleman (1966), have there been serious attempts to obtain measures of perceptions of environmental factors.

(d) Early emphasis on evaluation focused on individual achievement.

In more recent years the focus has been on the evaluation of group achievement to determine the effectiveness of teaching materials, instruction, and curriculum. The work of Rice (1897), Arnold (1916), Cronbach (1963), and Scriven (1967) testify to these changes in emphasis.

(e) With the expansion of educational involvement in the areas of the military, colleges and professional schools, and early childhood education, the need for an accompanying new evaluation concept has arisen.

Merwin (1969), in the sixty-eight NSSE yearbook states that changing concepts in evaluation have grown out of the technical development and the modes of interpretation which have developed to accompany new testing techniques. He showed that there are three major areas of concern.

1. The published Stanford Achievement Tests in 1923 by Terman, Ruch, and Kelly offered the first battery approach to testing across subject. This approach has been generally accepted as a source of achievement information for many years. The most prudent time to administer a test battery has been a point of controversy. School administrators have argued that the tests offer a measure of individual and group accomplishments and should be given at the end of a school year. Others have argued for fall testing to provide information to teachers as a basis for planning instruction.
2. When achievement tests were shown to be a more efficient and objective measure of achievement when compared to "essay" tests, the use of absolute (percentage) scores resulted in the development of a normative approach to testing. For several decades evaluation focused on the

development of instruments which reliably differentiate between individuals and interpreted the results of these instruments in terms of norms. Recently the focus has been to establish standards, as in the Oak Leaf Project at Pittsburgh (Glaser, 1968), which is a "mastery" testing. This type of testing is based on a child showing that he has accomplished a particular task or behavior to a certain degree of proficiency as required. Additional types of evaluation which have come from a competency--based on education are those which Burns (1972) speaks of: "When the method or way of performing (behaving) is important, a process measuring situation can be thought of as a test item. If the end result is more important than the method, a product measuring situation is required. Products can include plans, blueprints, drawings, paintings, tables, charts, diagrams, models, photographs, collections, specimens, stories, poems, and an infinite number of other real things. In many instances much can be inferred about a process from observing a product, the two are interrelated. Evaluations using processes and products are commonly more valid than merely testing at the verbal level, which may or may not indicate competence."

3. The interpretation of achievement in terms of potential has been used by educators for many years for identifying selected norm groups. Schudson (1972) has described one of these established norm groups as a "meritocracy." He states that through the use of College Boards to determine "admissions to certain selective colleges, an additional simultaneous choice is made in the selection of those individuals in a society who are to be the future rulers of that society and the holders of the wealth." The report of the Commission on Tests (1970) described the situation in the following manner: "Certainly it is particularly unfortunate that the characteristics that make for success in school work as it is commonly conducted are, if not specific to some segments of society, at least disproportionately distributed among its social classes and its racial and ethnic groups: Bowdoin College's admissions director, Richard Moll, told the press that the tests could not escape cultural bias and so 'tend to work in favor of the more advantaged elements of our society, while handicapping others.' Problems of interpretation have arisen when achievement scores have been regressed on aptitude scores giving 'expectancies.' A lack of understanding of the meaning of 'expectancy' has led to the ideas that 'underachievers' can come up to their predicted level of performance if they would just apply themselves, and an 'overachiever' is doing better than he is capable of doing. As a result of labelling children,

teachers when expecting low achievement will often get just what they expect, resulting in a phenomena which has been called the 'self-fulfilling prophecy.'"

A major consideration of educational evaluation in the beginning was the provision of information for the teacher's use in working with students. The resultant effect of the use of standardized tests in the early part of the century was a new potential for considering the outcomes of different groups on a common examination. The use of a common test to evaluate learning has spread from a schoolwide basis, to a statewide consideration, and currently to a national assessment.

Lewy (1973) has raised some serious questions concerning the use of achievement tests to discriminate both among individuals and among classes.

Item selection procedures which are recommended for constructing tests for individuals differentiation may not be adequate for tests for discrimination among classes. In spite of the practical difference between discrimination of these two types, educational research has not paid enough attention to the existence of such differences, and therefore little systematic study has been devoted to its implications for the planning of educational studies, for the construction of instruments, and for analyzing educational data.

Carver (1975) in reviewing the findings in the Coleman Report (Equality of Educational Opportunity Survey, 1966) pointed out the Coleman data was designed to be biased against finding significant educational effects for the same reasons cited by Lewy. He stated:

Given the impact of the Coleman Report on federal policy and the allocation of federal funds, it is important that the basis for such policy be on firm ground. It would be unfortunate if the data did not reflect what they were purported to reflect.

With the advent of district, state, and federal testing and the resultant use of these results to make decisions concerning the funding of educational projects, the necessity for continued research in evaluation to answer the problems cited has been mandated.

A review of the historical development of testing disclosed the necessity of developing reliable objective tests to measure student achievement. This need has continued and grown as the evaluation of learning has been used to research the effectiveness of certain curricula, instruction, and learning environments as well as to simply measure individual achievement. Based on the assumption that measures of evaluation should be objective, the technique in this study offers a means of evaluation which retains the well-established need for objective measures. In addition, the testing technique also emphasizes the measurement of individual growth and self assessment.

Historically Developed Criteria for Judging Evaluation Instruments and Measurement

The need for objective evaluation instruments and measurements has existed for a relatively long time, acting as an impetus for the development of criteria to determine whether or not any given instrument or measurement did what it was purported to do. These criteria will be used in Chapter V to help evaluate the study's testing technique.

The first of a series of publications designed to help test makers refine their instruments was Statistical Methods Applied in Education written by Harold Rugg in 1917. From Rugg's work came a series of criteria for judging the desirability of accepting a testing

instrument and its results. Gronlund (1971) lists and defines these criteria as validity, reliability, and usability.

Validity

Validity refers to the extent to which the results of an evaluation procedure serve the particular uses for which they are intended. Three types of validity have been identified and are now commonly used in educational and psychological measurement: (1) content validity, (2) criterion related validity, and (3) construct validity.

Gronlund has defined these concepts:

1. Content validity may be defined as the extent to which a test measures a representative sample of the subject-matter content and the behavioral changes under consideration.
2. Criterion-related validity may be defined as the extent to which test performance is related to some other valued measure of performance.
3. Construct validity may be defined as the extent to which test performance can be interpreted in terms of certain psychological constructs.

Gronlund has pointed out additional factors found in the test instrument which, if ignored, will lower the validity of the test results.

1. Unclear directions.
2. Reading vocabulary and sentence structure too difficult.
3. Inappropriate level of difficulty of test items.
4. Poorly constructed test items.
5. Ambiguity.
6. Test items inappropriate for the outcomes being measured.

7. Test too short.
8. Improper arrangement of items.
9. Identifiable pattern of answers.

Factors which influence validity that can be found in the administration and scoring of a test are the following:

1. Cheating.
2. Failure to follow directions.
3. Ignoring time limits.
4. Giving pupils unauthorized assistance.
5. Errors in scoring.
6. Poor physical environment.

Conditions that might adversely affect test validity which are due to personal factors are:

1. Motivation.
2. Anxiety.
3. Fatigue.
4. Illness.
5. Test-wiseness (ability to discern cues to correct responses from the test itself).
6. Response set (consistent tendency to follow a certain pattern in responding to test items).

Gronlund summarizes the nature of validity thus:

the validity of test results is based on the extent to which the behavior elicited in the testing situation is a true representation of the behavior being evaluated. Thus, anything in the construction or the administration of the test which causes the test results to be unrepresentative of the characteristics of the person tested contributes to lower validity. In a very real sense, then, it is the user of the test who must make the final judgment concerning the validity of the test results. He is the only one who knows how well the test fits his particular use, how well the testing conditions were controlled and how typical the responses were to the test situations.

Reliability

Reliability refers to the results obtained with an evaluation instrument and not to the instrument itself. According to Gronlund (1971),

Reliability refers to the consistency of measurement. That is, to how consistent test scores or other evaluation results are from one measurement to another. . . . A closely related point is that an estimate of reliability always refers to a particular type of consistency. Test scores are not reliable in general. They are reliable (or able to be generalized) over different periods of time, over different samples of questions, over different raters, and the like. It is possible for test scores to be consistent in one of these respects and not in another. The appropriate type of consistency in a particular case is dictated by the use to be made of the results. . . . Treating reliability as a general characteristic can only lead to erroneous interpretations.

Gronlund adds that reliability merely provides the consistency which makes validity possible. A highly reliable measure may have little or no validity.

Factors which may influence reliability are:

1. Length of test--In general, the longer the test the higher reliability.

2. Spread of scores--In general, the larger the spread of scores, the higher the estimate of reliability.
3. Difficulty of test--Tests which are too easy or too difficult for the group members taking it will tend to provide scores of low reliability.

Usability

Usability refers to the practical considerations of selecting an evaluation instrument. Some of these are:

1. Ease of administration.
2. Time required for administration.
3. Ease of scoring.
4. Ease of interpretation and application.
5. Availability of equivalent or comparable forms.
6. Cost.

Review of the Research in Math Instruction

The definition of a math lab contributed by Kerr (1974) identifies the areas of research to be reviewed if math labs can be thought of as effective environments for learning.

The mathematics laboratory is a strategy of instruction in which the learner himself interacts with mathematics and its real-world applications. The techniques used in a laboratory strategy may be varied; they may include discussion, discovery activities, model construction or even some directed teaching. Likewise the interaction of the learner with mathematics and its applications may vary. But the laboratory strategy focuses the learner's attention and activities on the relationship between mathematics and its real-world applications.

The real world applications of mathematics take the form of models which demonstrate the mathematical concepts in a meaningful manner to the learner. On the basis of the research evidence put forth by the 20 studies conducted to determine the effectiveness of using models and activity oriented classrooms in teaching mathematics in kindergarten through third grades, it does appear that the use of mathematical models and activities contributed to effective teaching. Table 1 presents a summary of these studies. Aurich (1963), Hollis (1964), Crowder (1965), Nasca (1966), Williams (1967), Howard (1969), and Wynrath (1970) found significance in favor of the experimental groups using models and activities. Weber (1969) did not find significance, but did find a trend favoring the use of manipulatives. Two additional studies, by Norman (1955) and Ekman (1966), did not find significance for either the control or experimental groups at the end of the instructional period, but did find the experimental group showed superior retention two weeks and three weeks, respectively, after the instructional period had ended. Only one of the 20 studies showed the "traditional" method of instruction produced significance in achievement. This study, conducted by Passy (1963), used Cuisenaire rods and offered the only evidence that a traditional approach can be more effective than teaching with models and activities.

From the research charted in Table 2, it seems apparent that using models does not hurt the learner's ability to comprehend mathematical concepts. Studies by Dawson and Ruddell (1955), Carmody (1970), Bisio (1970), and Nickel (1971) show significant results for the use of

Table 1. Summary of the effect of activity and model methodologies on the learning of mathematics in kindergarten through third grade

Author	Grade Level	Model	Test Used	Significant Difference In Favor Of	Mathematical Content
Norman (1955)	third	concrete and semiconcrete models	author constructed	neither group at the end of instruction; concrete and semi-concrete at the end of two weeks	division of whole numbers
Eidson (1956)	early elementary	many multisensory aids	standardized achievement	neither	arithmetic in lower grades
Sole (1957)	early elementary	manipulative aids	standardized achievement	neither	arithmetic in lower grades
Seick (1959)	second and third	multisensory aids	author constructed	neither	computation and arithmetic reasoning
Aurich (1963)	first	Cuisenaire rods	standardized achievement	Cuisenaire treatment	total range of first grade work
Haynes (1963)	third	Cuisenaire rods	author constructed	neither	multiplication
Passy (1963)	third	Cuisenaire rods	standardized achievement	traditional treatment	computation and arithmetic reasoning
Lucow (1963)	third	Cuisenaire rods	author constructed	neither	multiplication and division
Hollis (1964)	first	Cuisenaire rods	standardized achievement	Cuisenaire treatment	total range of first grade work
Crowder (1965)	first	Cuisenaire rods	standardized achievement	Cuisenaire treatment	total range of first grade work
Nasea (1966)	second	Cuisenaire rods	standardized achievement	Cuisenaire treatment	total range of second grade work
Lucas (1966)	first	Dienes arithmetic blocks	standardized achievement and author constructed	Dienes treatment for conservation of number and conceptualization of mathematical principles; traditional for computation and solving of verbal problems	identified in projection terms: multiplication of relations and addition-subtraction relations
Ekman (1966)	third	counters	author constructed	neither at end of instruction; concrete model group on a retention test	addition and subtraction algorithms
Weber (1969)	first	manipulative and concrete	standardized achievement and author constructed	neither but a trend favored manipulatives	total range of first through third grades
Howard (1969)	early elementary	concrete materials	author constructed	concrete materials	sorting, counting classifying and patterning sets
Wynrath (1970)	kindergarten	games	standardized achievement	games	total range of kindergarten and first grade work
Moody, Abell & Bausell (1971)	third	manipulative and concrete materials	standardized achievement	neither	multiplication
Ropes (1972)	second	multisensory aids	standardized achievement and author constructed	neither	total range of second grade work

Table 2. Summary of studies to determine the effectiveness of teaching with models and activities in grades four through six

Author	Grade Level	Models Used	Test Used	Significant Difference In Favor Of	Mathematical Content
Price (1950)	fifth and sixth	multisensory aids	author constructed	neither	division of fraction
Howard (1950)	fifth and sixth	concrete and semiconcrete	author constructed	neither at end of instruction; semi-concrete three months later	total range of fifth and sixth grade work
Dawson & Ruddell (1955)	fourth	many diverse models	author constructed	concrete-model group	division of whole numbers
Anderson (1957)	eighth	various visual tactile devices	author constructed	neither	area, volume and pythagorean theorem
Mott (1959)	fifth and sixth	many multi-sensory aids	standardized achievement	neither	measurement
Spross (1962)	fifth and sixth	concrete aids that had cultural significance	standardized achievement	neither	total range of fifth and sixth grade work
Trueblood (1967)	fifth and sixth	manipulation of aids and demonstration of aids	standardized achievement	demonstration of aids	fractions
Toney (1968)	fourth	manipulation of aids and demonstration of aids	standardized achievement	neither	fourth grade content
Green (1969)	fifth	diagrams cardboard sticks	standardized achievement	neither	multiplication of fractions
Carmody (1970)	sixth	concrete and semiconcrete	author constructed	concrete and semiconcrete	sixth grade work
Bisio (1970)	fifth	demonstrated manipulatives	author constructed	manipulatives	fractions
Wilkinson (1970)	sixth	laboratory materials	standardized achievement	neither	metric geometry
Nickel (1971)	fourth	abstract picture and diagrams; concrete	standardized achievement	multi-model approach	fourth grade work
Ropes (1972)	sixth	laboratory materials	standardized achievement	neither	sixth grade work

models in teaching. Howard (1950) showed that there was no significant difference between treatment groups until a test was administered three months later to make a determination on retention. On the retention test the group using the models did significantly better.

The summary of results shown in Table 3 appears to reverse the findings in the early elementary studies. Instruction using models is less effective than traditional approaches. This finding was borne out by the work of Johnson (1970), Cohen (1970), Schwartz (1971), and Shoecraft (1971). Low achievers showed a need for aids in instruction in the Shoecraft (1971) study by showing significant results in group achievement. Waslyk (1970) showed significant results for his experimental group when working with measurement concepts using concrete models.

In reviewing this research, several questions occurred to this reader concerning the wisdom of accepting many of the results as an accurate measure of the effectiveness of model and activity teaching. Two such reservations are noted below.

1. Key words and procedures in the study lacked operational definitions. Therefore, variables which might have affected the results remain undisclosed. This lack of definition also affects replicability.
2. Concepts taught at the concrete, pictorial-diagrammatic level of representation were primarily evaluated at the abstract level of representation. This cannot help but place the results of teaching which uses concrete and semiconcrete mathematical aids and models at a disadvantage.

Table 3. Summary of the effect of activity and model methodologies on the learning of mathematics in grades seven through twelve

Author	Grade Level	Model Used	Test Used	Significant Difference In Favor Of	Mathematical Content
Cohen (1959)	twelfth	physical representation of geometric forms	standardized achievement	neither	solid geometry
Eveid (1964)	junior high	many diverse aids	standardized achievement	neither	range of junior high work
Vance (1969)	seventh and eighth	many diverse aids and activities	standardized achievement	neither	seventh and eighth grade work
Waslyk (1970)	ninth	concrete	standardized achievement	concrete	measurement
Johnson (1970)	seventh	many	standardized achievement and author constructed	textbook oriented instruction	number theory, geometry, measurement and rational numbers
Cohen (1970)	middle school	many diverse aids	standardized achievement	traditional instruction	fractions
Schwartz (1971)	seventh and eighth	many diverse aids	project constructed	seventh grade, neither; eighth grade traditional	seventh and eighth grade work
Shoecraft (1971)	seventh and ninth	many diverse aids	standardized achievement	aids for low achievers; traditional for middle and high achievers	problem solving in algebra

Despite the criticisms which might be leveled at the research cited, the results certainly can be accepted as strong evidence in support of model and activity learning. In the majority of cases these studies still found significant results in favor of such learning, even though the instruments used to measure learning placed them at a disadvantage. These instruments measured learning using symbolic concept representation, whereas a child using models and activities experiences concrete or pictorial-diagrammatic concept representations.

If math labs which place an emphasis on model and activity learning are themselves to be more accurately evaluated in terms of their effectiveness in teaching math, it is necessary for new methods of evaluation to be devised which will incorporate the objective nature of standardized tests and offer a means of evaluating learning at the concrete and pictorial-diagrammatic representation of concepts. With this purpose in mind this study was undertaken.

Evaluation Methods in Assessing Learning in a Math Lab

In addition to standardized tests in evaluating learning in a math lab, a few other methods have been employed.

Anecdotal Records

Anecdotal records are the objective, as opposed to interpretive, descriptions of pupil behavior written by the teacher on a daily or frequent basis. Gronlund made the following suggestions concerning the keeping of these records:

1. Confine observations to those areas of behavior that cannot be evaluated by other means.
2. Limit observations of all pupils at any given time to just a few types of behavior.
3. Restrict the use of extensive observations of behavior to those few pupils who are most in need of special help.

Rating Scales

Rating scales provide a systematic procedure for obtaining and reporting the judgments of observers. A rating scale consists of a set of characteristics or qualities to be judged and some type of scale for indicating the degree to which each attribute is present. According to Gronlund, the rating scale is valuable only to the extent it is carefully prepared and appropriately used. It should be constructed in accordance with the learning outcomes to be evaluated, and its use should be confined to those areas where there is a sufficient opportunity to make the necessary observations. If these two principles are properly applied, a rating scale serves several important evaluative functions: (1) It directs observation toward specific and clearly defined aspects of behavior; (2) it provides a common frame of reference for comparing all pupils on the same set of characteristics; and it provides a convenient method for recording the judgment of the observers.

The following principles were listed in Gronlund as important characteristics to be considered in the preparation or selection of a rating scale:

1. Characteristics should be educationally significant.
2. Characteristics should be directly observable.
3. Characteristics and points on the scale should be clearly defined.
4. Between three and seven ratings should be provided and raters should be permitted to mark at intermediate points.
5. Raters should be instructed to omit ratings where they feel unqualified to judge.
6. Ratings from several observers should be combined, whenever possible.

Checklists

According to Gronlund,

A checklist is similar in appearance and use to the rating scale. A rating scale provides an opportunity to indicate the degree to which a characteristic is present or the frequency with which a behavior occurs. The checklist, on the other hand, calls for a simple "yes-no" judgment. It is basically a method of recording whether a characteristic is present or absent, or whether an action was taken or not taken. Checklists are especially useful in evaluating those performance skills that can be divided into a series of clearly defined, specific actions.

In summary, the major points to be considered in developing a checklist, according to Gronlund, are: (1) Identify and describe clearly each of the specific actions desired in the performance; (2) add to the list those actions which represent common errors, if they are limited in number and can be clearly identified; (3) arrange the desired actions and likely errors in the approximate order in which they are expected to occur; and (4) provide a simple procedure for numbering the actions in sequence or for checking each action as it occurs.

Interview

An interview is an evaluation situation in which an examiner faces a student and asks questions to which the student is expected to respond. Suydam (1974) suggested the following procedure for a mathematics evaluation: (1) Face the student with a problem; (2) let him find a solution, as he tells you what he is doing; and (3) challenge him, to elicit his highest level of understanding.

All of the methods cited in this chapter to evaluate learning in a math lab are very time consuming in their preparation, administration, or both. These methods do offer a teacher a means of evaluating learning using concrete and pictorial-diagrammatic representations of concepts. Teachers using interviews or anecdotal records are able to judge whether or not a child has understood a concept which has been presented concretely by observing the behavior of the child using the concrete model and either writing down what has been observed or by asking the child questions about his behavior and recording the questions and responses.

The methods of evaluation cited here have the inherent problem of being subjective. The ability accurately to observe, record, and pose meaningful questions to determine the depth of learning being observed is highly dependent on the talents of the teacher doing the evaluating. This subjectivity may well bring back into the educational scene the kind of criticism which historically was shown to be valid with respect to the accuracy of measurement.

It is apparent that with all the methods and instruments available to evaluate learning, additional means are needed which (1) can measure learning with the myriad of levels of learning present in any given math lab, (2) require only a short time to prepare, administer, and correct, and (3) offer objective measures. This study offers a beginning in the research needed to establish the effectiveness of a testing technique which can accomplish these three necessary tasks.

Thresholding

Methods of evaluating student learning vary, but there is an emphasis on achievement tests, which are used to determine a level of functioning with respect to a norm. These norms are determined by testing youngsters to be normed and ascertaining levels of expectancy for children of a particular age or grade. Buswell and John, in Manual of Directions for Use with Diagnostic Charts for Individual Difficulties in Fundamental Processes in Arithmetic, state:

A standardized test in arithmetic will indicate whether a pupil is doing satisfactory or unsatisfactory work for a given school grade. It enables the teacher to identify those pupils who need special attention. However, the marked limitation of such a test is that it does not tell why the pupil fails nor how he has made his errors.

Since these tests do not attempt to determine a student's level of functioning within an area of arithmetic or mathematics, additional types, called diagnostic or inventory tests, have been developed. Meyers (1959) pointed out that there were 37 achievement and 10 diagnostic tests available in the area of arithmetic. The latter have a varied format, with a portion of them offering a sequenced

test from simple to complex problems within a computational skill area. To determine the level of functioning within a diagnostic test of this kind, a threshold of functioning is ascertained by observing at what point in this test a child either begins making more errors than correct responses or stops answering questions.

This method of determining the functioning level of an individual has a history beginning in 1860 with Fechner, who was the chief precursor of experimental psychology. He published a voluminous treatise on "Psychophysics" entitled Elements der Psychophysik. Initially a physicist who sometimes published philosophical works under a pseudonym, Fechner, because of his interest in philosophy, may have abandoned physics and been attracted to psycho-physics when he suffered from a nervous breakdown. He wanted to demonstrate the identity of mind and matter which to him were two faces of the same reality, and either of which was apparent according to whether one took an internal or an external point of view. His background in physics made him denounce reasoning as a valid source of knowledge. Seeking a scientific foundation for his knowledge, he hoped to determine a quantitative relationship between a physical stimulus and resulting conscious sensation. In his search for the scientific laws governing psycho-physics he devised suitable methods of experimentation and statistical treatment of data.

In his search for the relationship between mind and body, Fechner had to measure as accurately as possible the different thresholds of his subject. Threshold and its Latin equivalent, *lemen*, mean,

essentially, a boundary separating the stimuli that elicit one response from the stimuli that elicit a different response. Thresholds must be repeatedly tested, for they vary due to the nature of the senses. Therefore, a threshold is always a statistical value; customarily, the lower threshold is defined as the value of the stimulus which evokes a positive response on 50 percent of the trials.

The threshold technique developed by Fechner is a method of serial exploration. It consists of "descending" and "ascending" series, each carried far enough to locate the momentary transition point or threshold from one response category to another.

Using Fechner's technique, Binet attempted to measure a total intelligence by measuring its individual aspects. Terman (1917) has noted:

It was this point of view which long controlled the work of Binet, who, like others, began by attempting to get at intelligence by measuring memory, attention, sense discrimination and other individual functions.

Terman adds:

The assumption that it is easier to measure a part, or one aspect of intelligence than all of it, is fallacious in that the parts are not separate parts and cannot be separated by any refinement of experiment. They are interwoven and intertwined. Each ramifies everywhere and appears in all other functions. Memory, for example, cannot be tested separately from the associative processes. After vainly trying to disentangle the various intellectual functions, Binet decided to test their combined functional capacity without any pretense of measuring the exact contribution of each to the total product. Intelligence tests have been successful just to the extent to which they have been guided by that aim.

Terman concluded: "The proof of the Binet method is the fact that it works so well."

The technique of determining a threshold for the functional level of a sense with any individual, which began in psycho-physics with Fechner, was used by Binet in his initial experiments with the measurement of intelligence. When his first efforts failed, he continued using this technique, assuming that measuring sense functioning in combination would not diminish the effectiveness of the technique.

With the establishment of this technique in determining intelligence, thresholding has been employed in diagnostic inventory testing to ascertain a level of functioning within an arithmetic operation. Based on the assumption that thresholding is valid in diagnostic testing, the proposed research will attempt to shortcut this technique by demonstrating a more efficient method of determining a level of performance within an arithmetic operation.

The Use of Ambiguous Stimuli in Testing

Ambiguous stimuli were first employed in the area of projective techniques for identifying emotional problems of individuals. By placing a stimulus, which could have many responses, before an individual, much was learned about the person's inner thoughts. Rorschach's inkblots projective approach was a precursor to a variety of projective techniques, including interpretation of drawing, painting, handwriting, stories, fantasies, play, and drama. Exner (1974) was noted:

Although Rorschach first became interested in the use of inkblots to study psychopathology about 1911, it is doubtful that he undertook any serious investigation of their usefulness until 1917. In that he died in 1922, he probably spent no more than between 3 and 4 years working intensively with them.

Before his death, Rorschach did offer a variety of postulates concerning specific test features, especially form, color, and human movement. He did not formulate a global theory of the test and was quite conservative in discussing its potential usefulness. After his death, five major systems or approaches in using the Rorschach developed. These five systems have caused much controversy in the use and interpretations of the instrument and its results. Despite all the controversy, Exner (1974) pointed out that 60 percent of all patients in a clinical situation in 1971 were administered the test.

Aside from measuring psychopathology with projective techniques, attitudes have also been measured using ambiguous visual stimuli. Alberts (Suydam 1974) has developed a test using 21 cartoon-like drawings. Children are asked to respond to these by associating themselves with the character portrayal.

Self-reports which request that a student relate what he has learned in a given class or with a given instructor are common examples of uses of an ambiguous verbal stimulus.

To test a person's mathematical creativity, Evans (Suydam 1974) has designed a test for late elementary and early junior high school students which presents an ambiguous math situation. The student is expected to respond in as many different ways as possible. Responses are scored with respect to number, number of different kinds, and degree of uncommonness.

The evaluation of academic achievement as employed in this study appears to be a new area for using ambiguous stimuli. But the technique has a long history in the field of psychological assessment, where the Rorschach and Thematic Aperception Test have been used for diagnostic purposes in mental health for more than half a century.

CHAPTER III

PROCEDURE AND METHODOLOGY

The setting, the sample, the examiners who used the proposed technique, and the instrument used for validating the technique are described in this chapter. In addition, the procedure for determining a child's ability accurately to assess and communicate what he knows about addition and subtraction using symbolic models of concept representation, as well as the methods of analyzing the collected data, will be discussed.

Setting and Sample

This study was conducted in the Muncie, Indiana, school system and at the laboratory school at Ball State University using 161 elementary students. The Muncie schools in the study are located in an area of mixed socioeconomic populations. The predominant races represented in Muncie are Negroid and Caucasian. Burris, the Ball State University laboratory school has a mixed cultural, racial, and economic population, and 30 percent of the students have learning disabilities. These children are channeled into the regular school classrooms.

The Muncie schools were selected in consultation with the office of the superintendent of schools and members of the administration who were familiar with the type of school populations. Schools

with the most diverse composition with respect to racial groups and economic levels were selected.

The testing procedure was administered both to groups of children and to individual children. At Burris children were grouped in classrooms with three grades in each class. All classrooms in the elementary portion of the school were either a 1-3 grade group or a 4-6 grade group. There were four classrooms of each grouping. Six Muncie classrooms in six different schools were chosen. There were two first grades, three second grades, and one fifth grade used. The entire classroom of children in the Muncie schools and the entire population of Burris youngsters in grades 1-6 were evaluated using the technique in the study.

Examiners

The examiners were both preservice and in-service teachers. The former came from the student body of Ball State University and were majoring in elementary or special education. Sections of college juniors and seniors taking methods classes and who were scheduled to tutor were asked to use the technique in this study to determine the level of development of their child or small group of children within an operation prior to tutoring for fall and winter quarter (1974-1975). The in-service teachers were from the Muncie school system. They were selected by their principals from the schools recommended by the Muncie school administration. All six teachers who were asked to participate accepted.

Instruments and Methods Used for Validating the Technique in This Study

Two sequenced tests were written for the study. A subject's level of functioning on either or both of these tests was determined by using Fechner's technique of thresholding. Another measure of the child's level of functioning was taken using the technique of this study. This level was determined by comparing the child's submitted problem to the level of the test designed for the study. The number of the level which most nearly corresponded to the submitted problem was then given to the submitted problem. This resulted in each child having two scores in the form of two level numbers--one from the sequenced achievement test prepared for the study, and one from the technique being researched.

A commercially prepared test, Fundamental Processes in Arithmetic, devised by Buswell and John and published by Bobbs-Merrill Company, Inc., was used as a guide for sequencing problem levels in the tests written for the study. A copy of the commercial test may be found in Appendix C. One additional problem per level was added to increase reliability, but no more than one was added in an effort to minimize test fatigue. Copies of the tests prepared for the study are found in Appendix C.

Procedure

The examiners were given a procedure sheet (Appendix B) explaining what they were to do. This sheet requested the following:

1. Ask the child to be evaluated to, "Show me the hardest problem that you have learned to do in _____ and write the answer." (The participating college students used the technique with all four operations in whole numbers and fractions, but only the addition and subtraction data were analyzed.)

2. If the child, when writing an addition problem, wrote one having all zeros except for one digit in each addend, for example, $1000 + 2000 = 3000$, then the examiner was to request that the child write a problem with no zeros, except possibly in the answer. (In the pilot studies, when children submitted memorized responses, the level of functioning was not discernible to the examiner. Sometimes the child, when giving a $(1000 + 1000 = 2000)$ response, indicated that he could only add a one-digit number to another one-digit number. In other instances, the problem indicated that he could add numbers in the thousands.)

3. After the child submitted his problem and answer, the test written for the study in addition or subtraction was given to him.

4. Last, the test was to be collected when the child wished to hand it in.

5. A request to fill out a data sheet concluded the directions on the procedure sheet.

To compile the data, two students from Ball State University, one in graduate school in elementary education and one a senior in secondary math education, determined the level of the submitted "hardest" problem by comparing the problem to the tests written

for this study in the appropriate operation and selecting the level that most corresponded to the submitted problem. This was done for all submitted problems first. Then the tests written for the study were scored using the thresholding technique to score the tests. The thresholding technique of scoring was used in the following way: When a child missed all three problems at a given level, his functioning level was determined to be at one level before the missed group of problems.

To test the following hypothesis, a criterion for determining high, average, and low achievers was established.

- B1 There will be no significant differences between the high, average, and low achievers as determined by the Iowa Achievement tests in their ability to assess their level of abstract achievement.

A child was judged to be a high achiever if his score on the Iowa Achievement test was in the 85th percentile or above, an average achiever if his score on the Iowa Achievement test was between the 30th and 85th percentile, and a low achiever if his score on the Iowa Achievement test was on the 30th percentile or below.

To test hypothesis B2, which reads as follows:

- B2 There will be no significant differences between the high, average, and low achievers as determined by teacher judgment, in their ability to assess their level of abstract achievement.

children were determined to be high, average, or low achievers simply on the basis of how a teacher viewed their achievement.

Hypothesis B5 states:

- B5 There will be no significant differences between children from high, average, and low family incomes in their ability to assess their level of abstract achievement.

To test this hypothesis, the following criteria to determine the category of family income which most nearly corresponded to each child was used: Scale of family incomes--high, over \$25,000; average, \$4,681 to \$24,999; and low, below \$4,681.

Methods of Analyzing Data

To establish a measure of validity with respect to the testing technique in this study, a comparison of results was made between the test written for this study, using the concept of thresholding to determine the level of functioning, and the technique in this study. The comparison took the form of a correlation which was hypothesis A of this study. It states:

- A There will be a high correlation between the results of testing using a diagnostic test and the results of testing using the technique being studied.

Constructing a scattergram on the results of the test written for this study together with the results of the technique in this study, a linear relationship was noted for both operations. (See

accompanying scattergram, Figure 1.) On the vertical axis of the scattergram are listed all possible levels (1 through 22) that a child could attain on the tests designed for the study. The horizontal axis lists levels 1 through 22, which are all the possible scores attainable by the testing technique in this study. Each pair of scores which a child acquires through testing are used as coordinates of points in the scattergram.

Since a linear relationship was apparent from the data, a decision to use the Pearson product-moment correlation coefficient was made. This correlation coefficient is denoted by r_{xy} . It can be expressed as the covariance of two variables, divided by the standard deviation of each of the variables:

$$r_{xy} = \frac{S_{xy}}{S_x S_y}.$$

The computational formula which was used is:

$$r_{xy} = \frac{n \sum X_i Y_i - (\sum X_i) (\sum Y_i)}{\sqrt{[n \sum X_i^2 - (\sum X_i)^2] [n \sum Y_i^2 - (\sum Y_i)^2]}},$$

where X and Y are the variables to be correlated, and n is the total number of subjects.

In an effort to test the following hypotheses it was necessary to establish a criterion for determining which children were successful in communicating their level of functioning by submitting a problem in addition or subtraction which they thought was the "hardest" that they could do.

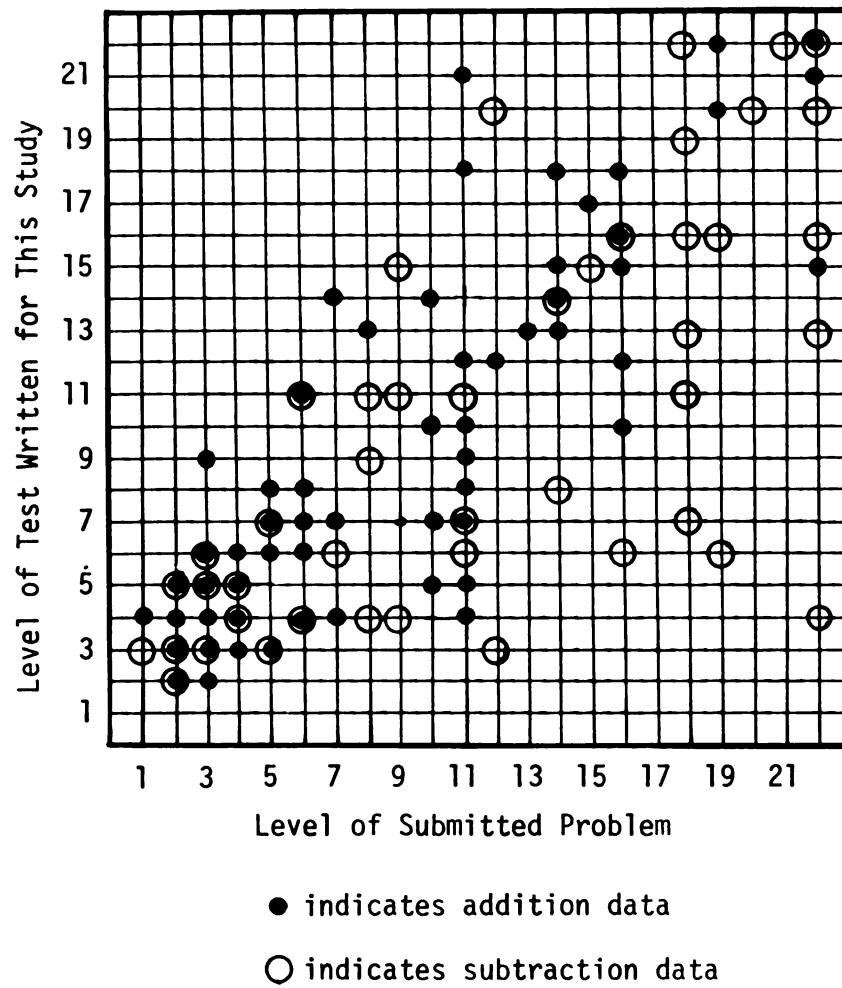


Figure 1. Scattergram of the results of the test written for this study and the technique of this study.

- B1 There will be no significant differences between the high, average, and low achievers as determined by the Iowa Achievement tests in their ability to assess their level of abstract achievement.
- B2 There will be no significant differences between the high, average, and low achievers as determined by teacher judgment in their ability to assess their level of abstract achievement.
- B3 There will be no significant differences between Blacks and Caucasians in their ability to assess their level of abstract achievement.
- B4 There will be no significant differences between girls and boys in their ability to assess their level of abstract achievement.
- B5 There will be no significant differences between children from high, average, and low income families in their ability to assess their level of abstract achievement.

The level of the problem submitted was compared with the results of the test written for this study, which the children took in the same testing session. The criterion for a successful self-assessment was established as follows: When a child submitted a problem which was within two levels above or two levels below the level of functioning established by the test written for the study, he was judged to be successful in his ability to assess himself. In tabulating the results, dichotomous data were collected, with a "1" being given to successful students and a "0" to unsuccessful students.

whether

to asse

check

varian

when

The 1

were.

As a precaution to the subsequent use of t-tests to determine whether or not there were differences in group means in their ability to assess themselves (hypotheses B1 through B5), an F-test was used to check sample variances. When the tests showed no differences in sample variances, the following two-tailed t-test was used:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{Sp \sqrt{1/n_1 + 1/n_2}}$$

\bar{X}_1 = mean of one group;

\bar{X}_2 = mean of second group;

n_1 = number of responses in first group; and

n_2 = number of responses in second group; and

where
$$Sp^2 = \frac{(n_1-1) S_1^2 + (n_2-1) S_2^2}{n_1 + n_2 - 2}, \text{ and}$$

Sp^2 = total population variance;

S_1^2 = variance of first group; and

S_2^2 = variance of second group.

The limits were:

Upper = $t_1 - \alpha/2$;

Lower = $t \alpha/2$;

d.f. = $n_1 + n_2 - 2$; and

$\alpha = .05$.

The assumptions which were made by using this test statistic were:

1. X_1 and X_2 are normally distributed;
2. homoscedasticity; and
3. samples were randomly selected and independent.

In the determination of a racial bias with respect to what a child evaluates as "hard," as suggested by hypothesis C1 (there will be no significant differences between racial groups in what they perceive as "hard"), the submitted problems were studied in an attempt to ascertain appropriate groupings for the analyses. If a submitted problem fitted into more than one category, then a tally mark was placed in all appropriate categories. The addition data were grouped in the following manner:

1. addition with regrouping;
2. addition without regrouping;
3. problems with three digits or less;
4. problems with more than three digits; and
5. problems with multiple addends (more than two).

Subtraction was grouped into the following categories:

1. subtraction with borrowing;
2. subtraction without borrowing;
3. problems with three digits or less; and
4. problems with more than three digits.

The nature of the data collected to test hypothesis C suggested that a series of chi-square tests be used with $\alpha = .05$. The following test statistic was used:

where

η_1

$$\chi^2 = \sum_{i=1}^k \frac{[n_i - np_i]^2}{np_i},$$

where n_i is the observed cell frequency, n is the sum of $n_1 + n_2 + \dots + n_k$, and p_i is the expected frequency.

CHAPTER IV

PRESENTATION AND ANALYSIS OF THE DATA

The results of this investigation using the procedures and data analysis described in Chapter III are presented in this chapter. A presentation of the data demonstrating the correlation between the technique in this study and that of the test written for this study will be given first. A discussion of the results of determining whether a child can assess himself by the criteria established in this research will follow. Finally, a presentation of the data showing the different groups' ability to use the testing technique in this study, cited as hypotheses in the preceding chapter, and the data used to determine whether or not a racial bias exists with respect to what a child considers "hard" will be discussed.

Correlation Between the Technique in This Study and the Test Written for This Study

The test and the children's submitted problems were collected as described in the procedure sheet in Appendix B. After the collection of these papers, a senior student in secondary math education and a graduate student in elementary education from Ball State University determined the level of the submitted "hardest" problem by comparing the problem to the test written for this study in the appropriate operation and selecting the level that most corresponded to the

submitted problem. This was done for all submitted problems, first. Then the tests written for this study and taken by the children were scored using Fechner's thresholding technique to determine the child's level of performance on the test. The thresholding technique of scoring a test was used in the following way: When a child missed all three problems at a given level, his functioning level was determined to be at one level before the missed group of problems.

Each child in this study, thus, has two scores--one from his submitted problem and one from the test designed for the study. A Pearson product-moment correlation was used to test hypothesis A.

- A There will be a high correlation between the results of testing a child by a diagnostic test and the testing technique being studied.

A value of $r = .85$ for addition and $r = .81$ for subtraction was computed. The results do show that a high correlation was found between the diagnostic test designed for the study and the testing technique in this study. Constructing confidence intervals for these two correlations ($P = .99$), ρ was found to be between .75 and .91 for addition and .66 and .90 for subtraction. Therefore, it can be concluded that the technique in this study gave results which correlated quite well with the results of the tests designed for this study for both operations.

Child's Ability to Assess Himself

The percentage of students who submitted problems within two levels above or below the level of functioning indicated by the diagnostic test was calculated to be 62 percent with addition and 57 percent with subtraction. A breakdown of the addition data shows that 33 of the 91 students were nonassessors by the criteria stated in Chapter III. It was not possible to assess two of the students in the study because they refused to submit a problem, stating that they could not think of one. The nonassessors could be broken down into the following categories: (1) submitted a problem incorrectly solved; (2) submitted a problem below (less difficult) the level of functioning as determined by the diagnostic test; and (3) submitted a problem above (more difficult) the level of functioning as determined by the diagnostic test.

Two students solved their submitted problem incorrectly, making errors that they also made on their test. Of the remaining nonassessors, 16 achieved a higher level score on the diagnostic test than their submitted problem indicated that they could do. Of these 16, 8 submitted problems which placed them in levels 1-12. All the 1-12 levels require little understanding of place value, and children could use their fingers to give a correct answer to the problems. Therefore, the 8 children who suggested by their submitted problems that they considered a one-digit number plus a one-digit number as the "hardest" problem that they could do, correctly answered many problems

by treating a multi-digit number as a series of one-digit number problems, that is, $435 + 362$ equals: five plus two, three plus six, and four plus three. This became apparent by observing the errors in the problems that they had missed. All of these children had the following type of error:

$$\begin{array}{r} 738 \\ + 436 \\ \hline 11614 \end{array}$$

It would appear that the submitted problem more accurately depicted their level of functioning.

Of the 16 students, 2 submitted problems without regrouping, and on their tests they indicated, by correctly working problems without regrouping, that they could regroup. Another 2 of the 16 could do multiple addend problems, but did not submit one. Four of the children submitted three or four-digit numbers with regrouping in their problem, but went on to solve the five-digit number problems with regrouping on their diagnostic test.

Of the 35 children who were not evaluated as self-assessors, 15 submitted problems which were on higher levels than they had scored on the diagnostic test. Of the 15, 4 appeared to suffer from test fatigue, boredom, or some other conditions which stopped the child from working all the problems up to the level of the submitted problem. Six of the students submitted problems which had many zeros, that is, $200 + 300 = 500$. This type of problem in the pilot studies preceding this investigation were shown to be an unreliable indicator of the level of functioning. The addition of a one-digit number to a two-digit

number was sequenced by the traditional diagnostic test written for this study as three levels above the addition of two two-digit numbers. By solving the one-digit problems and missing the addition of two two-digit problems, 5 youngsters indicated that the sequencing was incorrect for them.

Looking at the data obtained using the operation of subtraction, 29 children did not correctly assess their level of functioning as defined by the researcher in Chapter III. The nonassessors could be distributed into the following categories: (1) submitted problems with incorrect answers; (2) submitted a problem below (less difficult) the level of functioning indicated by the traditional diagnostic test written for this study; (3) submitted a problem above (more difficult) the level of functioning indicated by the traditional diagnostic test written for this study; and (4) had difficulty with the sequencing used in constructing the diagnostic test written for this study.

Five of the nonassessors wrote problems with incorrect answers, thereby giving no level of functioning. Another 9 students simply stopped answering test problems or missed problems with fewer digits and borrowing, which in earlier parts of the test they had answered correctly. It appears that test fatigue or lack of reinforcement may have influenced this behavior. These students submitted problems on a more difficult level than their diagnostic test indicated that they could do. Seven students submitted problems which were easier than they actually could do as determined by the diagnostic test.

In the sequencing provided by the test written for this study, levels containing problems with borrowing were intermixed with levels without. The emphasized criterion for adding a level in the test written for this study was the number of digits in a number, that is, a three-digit number with borrowing was considered more difficult than a four-digit number without borrowing. This emphasis in sequencing caused problems for some youngsters. A child who submitted a problem made up of three-digit numbers without borrowing would miss all borrowing problems at levels with smaller numbers, causing him to be judged a nonassessor. This was the case for 8 of the 29 nonassessors.

Analysis of the Data Concerning Hypotheses B1 through B5 of the Study

To test the following hypotheses of this study a series of t-tests were used:

- B1 There will be no significant differences between the high, average, and low achievers as determined by the Iowa Achievement tests in their ability to assess their level of abstract achievement.
- B2 There will be no significant differences between the high, average, and low achievers as determined by teacher judgment in their ability to assess their level of abstract achievement.
- B3 There will be no significant differences between Blacks and Caucasians in their ability to assess their level of abstract achievement.

- B4 There will be no significant differences between girls and boys in their ability to assess their level of abstract achievement.
- B5 There will be no significant differences between children from high, average, and low income families in their ability to assess their level of abstract achievement.

Several F-tests were run first in order to determine whether or not there were equal variances in the sample populations. The results of those tests are presented in Table 4.

The number of subjects used for the F-tests was 152; 9 students were omitted from the analysis because they either did not submit a problem or answered their problem incorrectly. In either case, it was impossible to determine a level of functioning from the use of the technique in this study. Using an α level of .05, no significant differences were found between the variances of the groups.

After collecting the data sheet handed out with the procedure sheet, it was noted that no teachers in the study evaluated a child in a different category of achievement than the category in which the child had been placed by the Iowa tests. Therefore, hypothesis B2 was not analyzed separately. Since no differences in variances were indicated by the F-tests, the following two-tailed t-test was used:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{1/n_1 + 1/n_2}}$$

Table 4. F-tests for determining the differences in variance of the groups in the study

	Variance	Number of Subjects in Group	F-Test Value	d.f.	Level	Significance
Economic Level:			1.167	8, 133	.05	none
high	.28 ^a	9				
average	.24 ^a	134				
low	.28	9				
Achievement Level:			1.217	8, 129	.05	none
high	.28 ^a	9				
average	.23 ^a	130				
low	.27	13				
Sex:			1.006	66, 84	.05	none
boys	.24	85				
girls	.25	67				
Race:			1.043	22, 129	.05	none
Black	.24	23				
White	.23	129				

^aIndicate groups used in F-test.

\bar{X}_1 = mean of one group;

\bar{X}_2 = mean of second group;

n_1 = number of responses in first group; and

n_2 = number of responses in second group;

where

$$S_p^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2},$$

S_p^2 = total population variance;

S_1^2 = variance of first group;

S_2^2 = variance of second group;

Upper = $t_{\alpha/2}$;

Lower = $-t_{\alpha/2}$;

d.f. = $n_1 + n_2 - 2$.

The results of the two-tailed t-tests are shown in Table 5. The two means (.44 and .60) for the high and average family income levels, respectively, were used in the t-test. No significant differences were found with an α level of .05. The t-statistic was .989, with 141 degrees of freedom. Therefore, it was concluded that there were no differences between the high, average, and low income family children in their ability to use the testing technique in this study.

Table 5. Group differences in their ability to use the testing technique in this study

	Group Mean	Number of Subjects in Groups	Variance of Sample Population	t-Test Value	d.f.	Level	Significance
Economic Level:			.49	.989	141	.05	none
high	.44 ^a	9					
average	.60 ^a	134					
low	.44	9					
Achievement Level:			.48	1.199	141	.05	none
high	.44 ^a	9					
average	.63 ^a	130					
low	.46	13					
Sex:			.38	.165	150	.05	none
boys	.60	85					
girls	.59	67					
Race:			.48	.555	150	.05	none
Black	.57	23					
White	.63	129					

^aIndicate the groups used in the t-test.

The means with the widest spread for ascertaining whether or not there was a difference between high, average, and low achievers in their ability to use the testing technique of this study were .44 and .63 (high and average, respectively). No significant differences were found with an α level of .05. The t-statistic computed was 1.199, with 141 degrees of freedom. It was concluded, therefore, that children who are high, average, or low achievers are all equally able to use the testing technique in this study.

To test whether or not boys and girls were equal in their ability to use the testing technique in this study, a t-test with an α level of .05 was used. A t-statistic of .165, with 150 degrees of freedom, was computed. No significant differences were found.

A t-test with an α level of .05 was used to determine whether or not there was a difference between Black and Caucasian children in their ability to use the testing technique in this study. The t-statistic was found to be .555, with 150 degrees of freedom. It was concluded that Black and Caucasian children were equally able to use the technique in this study. It would, therefore, appear from the data that all groups in the study are equally able to respond to the open question with a self-assessment which has a high degree of accuracy.

Analysis of the Data Concerning
Hypothesis C in the Study

The child-submitted addition problems were studied, and a decision was made to use the following categories as a basis for grouping to determine whether or not a racial bias exists with respect to what a child considered "hard." If a submitted problem fitted into more than one category, then a tally mark was placed in all the appropriate categories. The categories for addition and subtraction are given below.

Addition:

1. addition with regrouping;
2. addition without regrouping;
3. problems with three digits or less;
4. problems with more than three digits; and
5. problems with multiple addends.

Subtraction:

1. subtraction with borrowing;
2. subtraction without borrowing;
3. problems with three digits or less; and
4. problems with more than three digits.

If a child submitted the following problem in addition,
 $638 + 494 + 863 = \underline{\hspace{2cm}}$, then a tally mark would be placed in the following categories: addition with regrouping, problems with three digits or less, and problems with multiple addends.

A chi-square test was used to analyze each category. The results of these tests can be found in Table 6 (addition) and Table 7 (subtraction).

In the addition category, two children did not submit problems, two children incorrectly solved their problems, and one child was Chinese, a category not considered in this research. Omitting these subjects, 88 children were left to be used for testing hypothesis C with respect to addition. The chi-square values were very low and non-significant. The values ranged from .0004 to .3450. No cultural bias in addition was found with respect to what a child perceived as "hard."

Table 6. Summary of the results of the chi-square tests with addition

	Number of Subjects in Group	χ^2 Value	d.f.	α Level	Significance
No regrouping:		.0004	1	.05	none
Blacks	13				
Whites	75				
Regrouping:		.0009	1	.05	none
Blacks	13				
Whites	75				
Multiple addends:		3.4500	1	.05	none
Blacks	13				
Whites	75				
Three digits or less:		.0015	1	.05	none
Blacks	13				
Whites	75				
More than three digits:		.0207	1	.05	none
Blacks	13				
Whites	75				

In the subtraction category, five subjects incorrectly solved their submitted problems, thus limiting the number of subjects to 63 for the analysis. Very low nonsignificant values for chi-square were found, the values ranging from .0144 to .8900. It therefore was concluded that no racial bias was found with respect to what is considered "hard" by a child within the operation of subtraction.

Table 7. Summary of the results of the chi-square test with subtraction

	Number of Subjects in Group	χ^2 Value	d.f.	α Level	Significance
No borrowing:		.7830	1	.05	none
Blacks	9				
Whites	54				
Borrowing:		.8900	1	.05	none
Blacks	9				
Whites	54				
Three digits or less:		.0114	1	.05	none
Blacks	9				
Whites	54				
More than three digits:		.0160	1	.05	none
Blacks	9				
Whites	54				

CHAPTER V

SUMMARY, GENERALIZATIONS, AND IMPLICATIONS FOR FUTURE RESEARCH

The effectiveness of a testing technique which employs an ambiguous stimulus to ascertain a level of functioning within the operations of addition and subtraction was the primary question which this study attempted to explore. Historically developed criteria for evaluating testing instruments and measurements taken from Chapter II will be used to summarize and generalize the findings on the effectiveness of the technique in this study. A summary and the resultant generalizations concerning the data on a child's self-assessment as well as the different groups' ability to use the technique in this study will be presented. An additional analysis of the distribution of percentage of correct-response scores with respect to the technique in this study, which lent support to the conclusions concerning the effectiveness of this technique will be offered. A review of the stated purpose of this study and the implications for future research will conclude the chapter.

Criteria for Judging Testing Instruments and Measurements

Criteria for judging testing instruments and measurements cited in Chapter II will now be used to evaluate the testing technique in this

study. By comparing the results of the test designed for this study with the results of the new technique, a measure of criterion-related validity was made. A correlation of $r = .85$ for addition and $r = .81$ for subtraction was found. Using a confidence interval to examine the combined correlations, it can be assumed that with a probability of .99, the correlation between the results of the test designed for this study and the results of the technique of this study will be in the interval of $r = .72$ and $r = .90$ for both operations.

A testing instrument with content validity should ask questions covering all levels of representation for all concepts which the examiner deems necessary to an understanding of the area being tested. Since the technique in this study has the specific questions concerning content being posed and answered by the individuals being tested, the content validity is dependent on the examinee's ability to pose valid questions.

Does the testing technique measure a child's depth of understanding and reasoning ability, or does it measure a memorized or rote learned piece of information or rule? The construct validity of the test which comes from the testing techniques in this study has not been explored. To say that the construct validity of a test derived from the technique in this study is, in general, the same as a sequenced diagnostic test might not be true, for no research has been conducted to show this.

Looking at additional factors found in the instrument, which if ignored would lower validity, there are several which are minimized by the technique in this study.

1. unclear directions--The directions were tested in pilot studies, and few children in those studies indicated that they did not know what was being asked of them. Confused children either asked questions or did not respond.
2. reading vocabulary and sentence structure too difficult--No child is asked to read anything more than he, himself, writes. The directions for the test are read aloud by the examiner.
3. inappropriate level of difficulty of test items--The level of difficulty is judged by the examinee. From the data it appears that most children submit the "hardest" problem that they can do.
4. poorly constructed test items--The examinee writes what is understandable to him, and any poorly constructed items offer to the examiner information about the examinee's level of understanding.
5. ambiguity--The questions are posed and answered by the examinee, thus eliminating ambiguity of specific questions.
6. test items inappropriate for the outcomes being measured--The examinee, by posing his own question in an area designated by the examiner, minimizes this problem. By submitting inappropriate questions, information concerning the level of functioning of a child is still made available to the examiner.
7. test too short--By asking the examinee to submit the "hardest" problem that he can do, the necessity for a lengthy test was minimized. By correlating the results with a lengthy test, as was done in this study, the validity was, to a large measure, substantiated.

8. improper arrangement of items--Since the child submits only one problem per area to be measured, no arrangement of items is necessary.
9. identifiable pattern of answers--This category does not apply to the technique in this study.

Several comments can be made concerning factors which influence validity that can be found in the administration and scoring of a test.

1. cheating--Since each child submits his own problem and answer, cheating could be easily detected and minimized.
2. failure to follow directions--The only directions given are oral. Since there is only one direction, it is very easy for an examiner to clarify any misconceptions.
3. ignoring time limits--No time limits are imposed by the technique in this study.
4. giving pupils unauthorized assistance--This problem could apply to the technique in this study.
5. errors in scoring--Since there is only one problem per area, the number of errors is minimized. But each problem is unique. Therefore, no general answer sheet is available.
6. poor physical environment--A poor physical environment could effect the results of the technique in this study. But the time needed to complete this test is minimized, so the effects of the environment would be minimized.

Concerning conditions that might adversely affect test validity which are due to personal factors, the following may be noted:

1. motivation--Motivation would be increased, for children would be asked to show what they can do without being confronted with tasks that they cannot do.
2. anxiety--Anxiety would be minimized, for the child is asked only to demonstrate what he can do.
3. fatigue--The initial fatigue that the child has when entering the testing situation would remain with this technique, but any additional fatigue would be minimized due to the shortness of the testing period.
4. illness--Illness would still effect the child's ability to function, but its affects would be minimized due to the shortness of the testing period.
5. test-wiseness--This does not apply, since the child writes his own exam.
6. response-set--This does not apply, since the test is only one-problem-per-area long.

In conclusion, it appears that the test has good general validity using the criteria cited to make the judgment. Additional research should be done to establish the construct validity of the response which each examinee submits. Categories of responses, as with psychological testing using ambiguous stimuli, may offer different constructs.

The reliability of the testing technique in this study was measured, in part, when it was shown that two ways of measuring a level of functioning had a high correlation. This correlation indicates a consistency of response in a single testing situation. Several other factors which may influence reliability were pointed out in Chapter II.

1. The length of the test is a factor in reliability. Since the testing technique in this study requires only one problem per area for achievement evaluation, reliability might be questioned. The correlation data offer support to the reliability of the measurement along with the analysis of the percentage of problems correctly answered up to and including the level of the submitted problem.
2. Scores with a large spread are indicators of good reliability. The scores collected in this study have a very wide spread, as can be seen in Figure 2. (The horizontal axis of the figure lists the levels of the operations on the tests designed for the study. The vertical axis has a series of numbers from 1 through 22, which represents the number of students who submitted a problem. The coordinates of the points represent the level of the problem submitted and the number of students who submitted a problem at that level.
3. If a test is too easy or too difficult, the reliability of the results is threatened. The technique in the study asks that a child write a problem that he thinks is the hardest he can do.

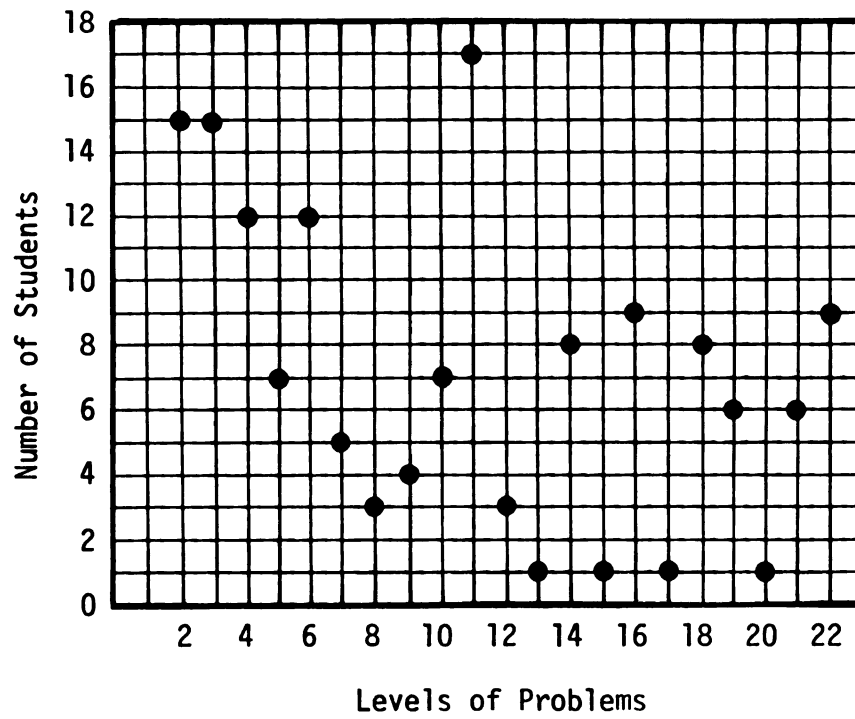


Figure 2. Spread of scores for addition and subtraction.

The data support that the child does just that. Therefore, it seems reasonable to assume that the test is neither too difficult nor too simple.

Usability is the last major factor to consider when making a decision about the advisability of using a particular test. The technique in this study has the following points in its favor: (1) It is easy to administer; (2) it requires a very short time to administer; (3) it is easy to score; (4) each child supplies equivalent forms of the test by identifying his level of performance with his own unique problem; and (5) little cost is involved.

The major problem that the technique in this study poses is one of interpretation of the results. If operations are tested using whole numbers and fractions, the problem is simplified. Materials are available which offer a sequencing of the skills involved in solving problems in these areas. But if the testing technique is to be used in other areas, analyses of what a child most likely knows in order to pose and answer a question in the chosen area will have to be done in order to interpret the results.

Accuracy of a Child's Self-Assessment

Of the children in the study, 60 percent, according to the criteria established in Chapter III, could assess their level of functioning. In analyzing the 64 youngsters who were categorized as nonassessors, 40 of these may well have assessed themselves.

These children met the following problems with the criteria established for assessment:

1. Eight youngsters indicated that they regarded a one-digit plus a one-digit number as the "hardest" problem that they could do. On the test written for the study, they treated several multi-digit problems with the algorithm they claimed to know for one-digit addition and solved the problems correctly. From their errors on the test, the algorithm used was made apparent. Therefore, it appears that these eight children did indicate their level of functioning.
2. Thirteen youngsters submitted problems which were more difficult than they completed correctly on the test written for the study. These children either quit solving problems or made errors that they had indicated earlier in the testing situation were within their scope of knowledge. For example,

$$\begin{array}{rclcl}
 \begin{array}{r} 23 \\ +47 \\ \hline 70 \end{array} & \text{later} & \begin{array}{r} 234 \\ +478 \\ \hline 61012 \end{array} & \text{still later} & \begin{array}{r} 2359 \\ +6874 \\ \hline 9233 \end{array}
 \end{array}$$

It appears that these youngsters may well have indicated their level of functioning, but were judged as nonassessors because of test fatigue, boredom, or some other similar problem.

3. Thirteen of the children appeared to have problems with the way the test was sequenced. They submitted problems which were considered easier or more difficult than the problem which they answered on the test designed for the study. The discrepancy proved to be enough to have them evaluated as nonassessors.

4. Six children submitted problems with zeros despite the attempt by a specific direction on the procedure sheet to negate the possibility of this happening. More care should be taken to avoid this type of error in the administration of the testing technique. With proper questioning, these children may well have assessed themselves correctly.

In considering the additional data just cited, apart from the criteria cited for successful assessing, it is questionable whether the 40 children just reviewed really could not assess their level of functioning.

It would appear that for the children who seemed unable to use the technique in this study several procedural considerations might be noted:

1. Some children in the study refused to submit a problem because they could not think of a "hard" one, for all problems within the operation being tested were considered simple by them. An examiner may, when noting the absence of a response caused by the cited difficulty, encourage a child to relate the fact in writing that all problems seem simple, thereby encouraging an honesty of response and a possible accurate assessment.
2. If a child were to submit more than one "hard" problem, an incorrect response may be more accurately evaluated by observing whether the error occurs again or whether it is a simple "foolish" inaccuracy.

Different Groups' Ability to Use the Testing Technique in This Study

Examining the data in the study concerning the different groups' ability to assess themselves (boys-girls, high-average-low achievers, children from high-average-low income families, and Blacks-Whites), all groups were shown to be able to use the testing technique equally effectively. The thought of using a test which has no built-in advantages or disadvantages for those children who in the past have suffered unfair discrimination from evaluation methods is very exciting. The possible use of the technique in this study to measure achievement in other content areas or even in intelligence testing may well offer a solution to the biased results present in testing today.

A Racial Bias with Respect to What Is "Hard"

No bias was found among Black and White children with respect to what is considered hard within the operations of addition and subtraction. Additional investigations may find biases where operations or realms of numbers are more complex.

Analysis of the Distribution of Percentage of Correct Response Scores with Respect to the Technique in This Study

When a child submits a problem as the "hardest" that he can do, can it be assumed that the levels considered simpler or less difficult are mastered? Using the sequencing of the test designed for this study and identifying the levels of this test to which the problem best

corresponds, an analysis of the percentage of correct responses was made. All those problems correctly answered up to and including the problem on the level of the submitted one were counted, and the percentage of correct responses was calculated. For addition, the mean was .86, with a standard deviation of .21 and a variance of .04. The subtraction data had a mean of .84, with a standard deviation of .18 and variance of .03. The data show that when two-thirds of a group of children submit a problem as the "hardest" one that they can do, they have mastered at least 65 percent of those problems sequenced as simpler and may have 100 percent of the simpler problems mastered.

Examining the percentage of problems answered correctly five levels above (more difficult) the submitted problem, the mean, variance, and standard deviation for addition were .21, .08, and .28, respectively. For subtraction, a mean of .20 with a variance of .10 and a standard deviation of .31 was found. It appears from the data that 68 percent of a group of children when submitting a "hardest-they-can-do" problem are able to work about one in five of a series of problems sequenced as more difficult.

A Review of the Stated Purpose of This Study

The purposes of this study were stated in Chapter I. How well these purposes were met will now be discussed.

1. The validation of the testing technique of this study has been, to a large measure, accomplished. Both the correlation and additional analyses concerning how well children individually and in groups can use this technique have yielded encouraging results.

2. The time required to prepare, administer, and correct the test in this technique is, indeed, minimized. The time required to think of the areas which need to be assessed and, possibly, to list them, is all the time required to use this technique. The administration and correcting time is also shortened, because the test itself is very short (one problem per area).

3. The shortness of the testing procedure directly affects the time that the student must spend in having his achievement evaluated.

4. The technique of this study indeed offers, on a daily basis, a collection of individual evaluations which will show the changes in what a child perceives as "hard" in his daily learning environment. If his environment has manipulatives or models, he can offer a problem which he can solve using these. Either he or his teacher can note on his paper what was used to help solve the problem.

5. The testing technique in this study places an emphasis on the examinee's ability to recognize what he can do. Through the repeated use of this technique a child may well be able to improve his ability to recognize self-growth; then, with guidance, he might be able to recognize what fosters self-growth and what deters it. With the emphasis on assessing what an individual knows instead of what he does not know, a testing situation will pose less threat to feelings of self-worth. With evaluation being done in terms of individual growth, the threat of having to meet group goals is also minimized. Both of these factors enhance the development of a good self-concept.

Implications for Future Research

The research proposed falls into two categories. The first is research on the usability of this technique in other areas besides the symbolic representation of addition and subtraction. The areas in mathematics education which might be researched using the technique of this study is the second category.

Usability of the Technique in This Study in Other Areas

Does an individual have the ability to recognize the knowledge and skills which he possesses? Can he relate what they are? These questions were answered in the affirmative with respect to the skill areas researched in this study. Studies to determine the effectiveness of evaluating learning with other operations, such as multiplication, and with different realms of numbers are also needed.

This research dealt primarily with measuring the level of skill development in computation. Can this technique measure concept learning? If college students were asked to note for themselves all the concepts that they felt had been presented to them in a given lecture, textbook chapter, laboratory manual, and so forth, could they then write the "hardest" question that they could think of which would test the understanding of each concept? By so doing, could a professor discern the degree of learning which has taken place for the student?

The greatest need for evaluative instruments and techniques is at the concrete and pictorial-diagrammatic representation of concepts and skills. With the encouraging results of this study using

symbolic representation, additional research is now called for using the technique in the evaluation of concept learning using other representations.

If a child cannot assess his knowledge initially, can he learn to do this? If he can assess himself and communicate his knowledge fairly well, can this skill be developed to a high degree of accuracy and broadened to include most of his learning experiences? Does the skill in self-assessment increase with the number of times that it is done? If a child cannot assess himself, can he be taught to do this? These are many of the questions which must be answered if the technique researched here is to be used with maximum understanding of its effects upon the examinee.

Areas of Mathematics Education to Be Researched Using the Technique in This Study

The technique in this study may prove fruitful in researching (1) the sequencing of mathematical models for the development of an understanding of a concept, (2) the carefully ordered presentation of concepts in learning a general area of mathematics, and (3) the effective ordering of the attributes of a concept for maximum clarification. Research will also have to determine whether there is a general sequencing of models, concepts, and attributes, or whether the orderings must take into account the background of each learner who will use them.

The effectiveness of different mathematical models for teaching concepts might also be explored with the technique in this study. In the pilot studies, children appeared to select a "hard" problem on the basis of the mathematical model that they were using at the time; that is, multiple addend problems were frequently submitted by children using Chip Trading to learn addition. Large numbers and problems with regrouping are very simply added using Chip Trading, but addition with several addends causes some problems. Studies to varify or negate the relationship between "hard" problems and models may prove valuable. When the most effective model is used to teach a particular concept to a child who finds the model readily understandable, learning would be greatly facilitated.

In general, the technique in this study offers a researcher the opportunity to collect evaluation data on a daily basis because of the simplicity of administration and the small amount of time required to complete the testing task. The daily evaluations make available information on the order in which skills and concepts are learned.

The examination of nonassessors' test papers indicated that some of these children found the sequencing of the test written for the study incorrect for them. They learned how to correctly answer levels on the test which were considered more difficult than the ones that they had missed. Another group of children seemed to agree with the sequencing by missing all the problems beyond a particular level. These data raised the issue of whether a sequence of learning

tasks could be written whereby all children would find the sequence correct for them, or whether the sequencing of learning tasks for individuals requires that the learner's background be taken into account. Since the testing technique in this study pointed out this discrepancy, it may be a useful tool to help answer the sequencing questions.

If the question used in the testing technique were altered to read: "Write a "hard" problem in _____ that you cannot answer" (the area to be evaluated would be read in the blank), the child would have to know enough about the area being evaluated to write a question, but not enough to answer it. This may well prove to be a way of ascertaining an appropriate "next" learning experience which would enable a child to solve his posed problem.

APPENDIX A

QUESTIONS TESTED IN PILOTS

APPENDIX A

QUESTIONS USED IN PILOTS

These questions are listed in order of greatest number of positive responses. If a child could not think of a response to the question, this was noted by the examiner. The question with the fewest number of "no responses" was selected to be used for the study.

1. Show me the hardest problem that you have learned to do in addition (subtraction, multiplication, or any other realm of study about which the examiner wishes to gain information) and write the answer.
2. Make up the hardest problem that you can in addition (subtraction, multiplication, and so forth). Solve it and write the answer.
3. Write the two hard problems in addition (subtraction, multiplication, and so forth) that we can put on ditto for the class to solve. Please include the answer.
4. Write down a problem that you can do in addition (subtraction, multiplication, and so forth), but maybe no one else can, and solve it.
5. Write a hard, tricky problem that only you can find the answer to.

Question 1 was amended to meet different assessment needs. If the question were used to measure a daily growth learning situation, it was worded: "Show me the hardest problem that you learned to do today and write your answer."

If a concrete or diagrammatic mode was being assessed, the question became: "Show me the hardest problem that you learned to do today and use the aid that you were working with to check your answer."

APPENDIX B

PROCEDURE HANDOUT

APPENDIX B

PROCEDURE SHEET

I wish to thank you for helping to collect data which will be used to determine the effectiveness of this testing technique.

1. Select the child or group of children that you wish to test.
2. Read to the child or group the following question, substituting the correct operation or area of mathematics that you would like them to consider when answering the question. I have used addition in the wording of this sample question. "Show me the hardest problem that you have learned to do in addition and write the answer."
3. When testing the area of addition, only, do the following: See if a child submits a problem with all addends using zeros except for the first digit. If he does, request that he write another problem with no zeros except for a possible zero in the answer.
4. When the child indicates that the task is completed, collect the problem. There is no time limit.
5. Pass out the diagnostic test appropriate to the area you are testing.
6. Ask the child to complete as many of the problems as he can, letting him know that there is no time limit.
7. Collect the diagnostic test.
8. Fill out the accompanying data sheet on the child.

Data Sheet

_____	_____
Child's Name	Age

Sex	

Achievement level as measured by the last Iowa Test child has taken:

Circle one: high average low

Economic level:

Circle one: (Over \$25,000) (\$4,681-\$24,999) (Below \$4,681)
 high average low

Race:

Circle one: Negroid Caucasian Other

Achievement level as measured by the child's classroom teacher:

 high average low

APPENDIX C

TESTS

Prod. No. 77856

PUPIL'S WORK SHEET
Diagnostic Chart for Fundamental Processes in Arithmetic
 Prepared by G. T. Burwell and Lenore John

THE TEST DIVISION OF
Bm The Bobbs-Merrill Company, Inc.
 4300 W. 62nd St. / Indianapolis, Indiana 46206

Printed in U. S. A.

ADD: School _____		Name _____	
(1) $\begin{array}{r} 5 \\ 2 \\ \hline \end{array}$ $\begin{array}{r} 6 \\ 3 \\ \hline \end{array}$	(2) $\begin{array}{r} 2 \\ 9 \\ \hline \end{array}$ $\begin{array}{r} 8 \\ 4 \\ \hline \end{array}$	(3) $\begin{array}{r} 12 \\ 2 \\ \hline \end{array}$ $\begin{array}{r} 13 \\ 5 \\ \hline \end{array}$	
(4) $\begin{array}{r} 19 \\ 2 \\ \hline \end{array}$ $\begin{array}{r} 17 \\ 9 \\ \hline \end{array}$	(5) $6 + 2 =$ $3 + 4 =$	(6) $\begin{array}{r} 52 \\ 13 \\ \hline \end{array}$ $\begin{array}{r} 40 \\ 39 \\ \hline \end{array}$	
(7) $\begin{array}{r} 78 \\ 71 \\ \hline \end{array}$ $\begin{array}{r} 46 \\ 92 \\ \hline \end{array}$	(8) $\begin{array}{r} 3 \\ 5 \\ 8 \\ 2 \\ \hline \end{array}$ $\begin{array}{r} 8 \\ 7 \\ 9 \\ 7 \\ \hline \end{array}$	(9) $\begin{array}{r} 53 \\ 8 \\ \hline \end{array}$ $\begin{array}{r} 7 \\ 89 \\ \hline \end{array}$	
(10) $2 + 5 + 1 + 8 =$ $4 + 9 + 4 + 6 =$	(11) $\begin{array}{r} 664 \\ 203 \\ \hline \end{array}$ $\begin{array}{r} 145 \\ 652 \\ \hline \end{array}$	(12) $\begin{array}{r} 35 \\ 234 \\ \hline \end{array}$ $\begin{array}{r} 601 \\ 78 \\ \hline \end{array}$	
(13) $\begin{array}{r} 69 \\ 12 \\ \hline \end{array}$ $\begin{array}{r} 38 \\ 84 \\ \hline \end{array}$	(14) $\begin{array}{r} 532 \\ 87 \\ \hline \end{array}$ $\begin{array}{r} 82 \\ 896 \\ \hline \end{array}$	(15) $\begin{array}{r} 13 \\ 7 \\ 5 \\ 2 \\ \hline \end{array}$ $\begin{array}{r} 8 \\ 9 \\ 33 \\ 8 \\ \hline \end{array}$	
(16) $\begin{array}{r} 268 \\ 961 \\ \hline \end{array}$ $\begin{array}{r} 943 \\ 128 \\ \hline \end{array}$	(17) $\begin{array}{r} 283 \\ 748 \\ \hline \end{array}$ $\begin{array}{r} 495 \\ 778 \\ \hline \end{array}$	(18) $\begin{array}{r} 34 \\ 33 \\ 55 \\ 94 \\ \hline \end{array}$ $\begin{array}{r} 66 \\ 98 \\ 68 \\ 49 \\ \hline \end{array}$	
(19) $\begin{array}{r} 13 \\ 587 \\ 46 \\ 131 \\ \hline \end{array}$ $\begin{array}{r} 66 \\ 989 \\ 896 \\ 467 \\ \hline \end{array}$	(20) $\begin{array}{r} 9361825 \\ 8758785 \\ \hline \end{array}$ $\begin{array}{r} 3907598 \\ 785763 \\ \hline \end{array}$	(21) $\begin{array}{r} 1 \\ 6 \\ 8 \\ 1 \\ 3 \\ 0 \\ 7 \\ 1 \\ 8 \\ 4 \\ 0 \\ 2 \\ 2 \\ \hline \end{array}$ $\begin{array}{r} 6 \\ 2 \\ 7 \\ 9 \\ 4 \\ 9 \\ 8 \\ 6 \\ 6 \\ 9 \\ 8 \\ 4 \\ 3 \\ \hline \end{array}$	
(22) $\begin{array}{r} 879 \\ 266 \\ 498 \\ 167 \\ 137 \\ \hline \end{array}$ $\begin{array}{r} 866 \\ 969 \\ 986 \\ 898 \\ 449 \\ \hline \end{array}$	(23) $\begin{array}{r} 817 \\ 7053 \\ 42610 \\ 92 \\ 938512 \\ \hline \end{array}$ $\begin{array}{r} 5134 \\ 73045 \\ 3 \\ 227528 \\ 242 \\ \hline \end{array}$		

SUBTRACT:

(1)	<div>5 3 —</div>	<div>8 8 —</div>	(2)	<div>7 − 1 = 9 − 0 =</div>	(3)	<div>19 2 —</div>	<div>15 4 —</div>	
(4)	<div>58 4 —</div>	<div>79 3 —</div>	(5)	<div>36 21 —</div>	<div>79 24 —</div>	(6)	<div>12 6 —</div>	<div>10 2 —</div>
(7)	<div>15 13 —</div>	<div>19 12 —</div>	(8)	<div>59 − 2 = 86 − 4 =</div>	(9)	<div>346 215 —</div>	<div>836 302 —</div>	
(10)	<div>189 45 —</div>	<div>399 70 —</div>	(11)	<div>61 2 —</div>	<div>75 9 —</div>	(12)	<div>56 48 —</div>	<div>42 36 —</div>
(13)	<div>92 64 —</div>	<div>42 19 —</div>	(14)	<div>528 64 —</div>	<div>292 94 —</div>	(15)	<div>1067 237 —</div>	<div>4498 825 —</div>
(16)	<div>624 193 —</div>	<div>852 308 —</div>	(17)	<div>431 162 —</div>	<div>963 594 —</div>	(18)	<div>950 376 —</div>	<div>507 221 —</div>
(19)	<div>9546 8687 —</div>	<div>9653 2954 —</div>	(20)	<div>5941 968 —</div>	<div>6805 978 —</div>	(21)	<div>132428 38679 —</div>	<div>823533 245838 —</div>
(22)	<div>10000 8192 —</div>	<div>80030 46759 —</div>						

Addition Test

Level 1	$\begin{array}{r} 2 \\ +1 \\ \hline \end{array}$	$\begin{array}{r} 5 \\ +1 \\ \hline \end{array}$	$\begin{array}{r} 7 \\ +1 \\ \hline \end{array}$
Level 2	$\begin{array}{r} 3 \\ +2 \\ \hline \end{array}$	$\begin{array}{r} 4 \\ +5 \\ \hline \end{array}$	$\begin{array}{r} 5 \\ +3 \\ \hline \end{array}$
Level 3	$\begin{array}{r} 8 \\ +3 \\ \hline \end{array}$	$\begin{array}{r} 6 \\ +7 \\ \hline \end{array}$	$\begin{array}{r} 2 \\ +9 \\ \hline \end{array}$
Level 4	$\begin{array}{r} 12 \\ +2 \\ \hline \end{array}$	$\begin{array}{r} 3 \\ +16 \\ \hline \end{array}$	$\begin{array}{r} 5 \\ +14 \\ \hline \end{array}$
Level 5	$\begin{array}{r} 19 \\ +6 \\ \hline \end{array}$	$\begin{array}{r} 15 \\ +8 \\ \hline \end{array}$	$\begin{array}{r} 9 \\ +14 \\ \hline \end{array}$
Level 6	$\begin{array}{r} 20 \\ +30 \\ \hline \end{array}$	$\begin{array}{r} 50 \\ +30 \\ \hline \end{array}$	$\begin{array}{r} 70 \\ +20 \\ \hline \end{array}$
Level 7	$\begin{array}{r} 23 \\ +36 \\ \hline \end{array}$	$\begin{array}{r} 34 \\ +52 \\ \hline \end{array}$	$\begin{array}{r} 57 \\ +22 \\ \hline \end{array}$
Level 8	$\begin{array}{r} 68 \\ +61 \\ \hline \end{array}$	$\begin{array}{r} 36 \\ +82 \\ \hline \end{array}$	$\begin{array}{r} 41 \\ +97 \\ \hline \end{array}$
Level 9	$\begin{array}{r} 2 \\ 4 \\ 7 \\ +1 \\ \hline \end{array}$	$\begin{array}{r} 6 \\ 8 \\ 5 \\ +7 \\ \hline \end{array}$	$\begin{array}{r} 9 \\ 3 \\ 7 \\ +3 \\ \hline \end{array}$
Level 10	$\begin{array}{r} 69 \\ +7 \\ \hline \end{array}$	$\begin{array}{r} 8 \\ +84 \\ \hline \end{array}$	$\begin{array}{r} 6 \\ +74 \\ \hline \end{array}$

Addition Test (continued)

Level 11	537 <u>+122</u>	603 <u>+115</u>	232 <u>+145</u>
Level 12	35 <u>+343</u>	67 <u>+112</u>	231 <u>+64</u>
Level 13	33 <u>+28</u>	42 <u>+58</u>	75 <u>+26</u>
Level 14	532 <u>+87</u>	94 <u>+937</u>	643 <u>+97</u>
Level 15	17 6 3 <u>+8</u>	4 2 27 <u>+19</u>	16 4 7 <u>+5</u>
Level 16	349 <u>+868</u>	914 <u>+879</u>	406 <u>+798</u>
Level 17	64 38 96 <u>+41</u>	17 33 14 <u>+73</u>	21 16 38 <u>+97</u>
Level 18	12 466 83 <u>+106</u>	343 8 14 <u>+173</u>	684 16 9 <u>+352</u>
Level 19	9416772 <u>+6541334</u>	7634215 <u>+4556148</u>	3716482 <u>+9784601</u>

Addition Test (continued)

Level 20

$$\begin{array}{r} 1 \\ 4 \\ 0 \\ 8 \\ 7 \\ 9 \\ 6 \\ 5 \\ 3 \\ 6 \\ 5 \\ 0 \\ + 1 \\ \hline \end{array}$$

$$\begin{array}{r} 2 \\ 1 \\ 9 \\ 3 \\ 4 \\ 0 \\ 6 \\ 0 \\ 7 \\ 4 \\ 2 \\ 0 \\ + 4 \\ \hline \end{array}$$

$$\begin{array}{r} 1 \\ 4 \\ 5 \\ 4 \\ 2 \\ 0 \\ 4 \\ 2 \\ 7 \\ 8 \\ 3 \\ 0 \\ + 3 \\ \hline \end{array}$$

Level 21

$$\begin{array}{r} 688 \\ 964 \\ 235 \\ 874 \\ + 118 \\ \hline \end{array}$$

$$\begin{array}{r} 603 \\ 715 \\ 404 \\ 670 \\ + 841 \\ \hline \end{array}$$

$$\begin{array}{r} 732 \\ 804 \\ 211 \\ 405 \\ + 607 \\ \hline \end{array}$$

Level 22

$$\begin{array}{r} 1 \\ 816 \\ 961453 \\ 4105 \\ + 63 \\ \hline \end{array}$$

$$\begin{array}{r} 816 \\ 37 \\ 9 \\ 4864 \\ + 718611 \\ \hline \end{array}$$

$$\begin{array}{r} 37 \\ 106 \\ 816439 \\ 4 \\ + 797 \\ \hline \end{array}$$

Subtraction Test

Level 1	$\begin{array}{r} 5 \\ -1 \\ \hline \end{array}$	$\begin{array}{r} 1 \\ -1 \\ \hline \end{array}$	$\begin{array}{r} 9 \\ -1 \\ \hline \end{array}$
Level 2	$\begin{array}{r} 6 \\ -3 \\ \hline \end{array}$	$\begin{array}{r} 5 \\ -4 \\ \hline \end{array}$	$\begin{array}{r} 8 \\ -6 \\ \hline \end{array}$
Level 3	$\begin{array}{r} 16 \\ -1 \\ \hline \end{array}$	$\begin{array}{r} 15 \\ -2 \\ \hline \end{array}$	$\begin{array}{r} 18 \\ -1 \\ \hline \end{array}$
Level 4	$\begin{array}{r} 16 \\ -5 \\ \hline \end{array}$	$\begin{array}{r} 15 \\ -4 \\ \hline \end{array}$	$\begin{array}{r} 18 \\ -6 \\ \hline \end{array}$
Level 5	$\begin{array}{r} 48 \\ -2 \\ \hline \end{array}$	$\begin{array}{r} 78 \\ -4 \\ \hline \end{array}$	$\begin{array}{r} 63 \\ -2 \\ \hline \end{array}$
Level 6	$\begin{array}{r} 28 \\ -17 \\ \hline \end{array}$	$\begin{array}{r} 49 \\ -22 \\ \hline \end{array}$	$\begin{array}{r} 97 \\ -35 \\ \hline \end{array}$
Level 7	$\begin{array}{r} 13 \\ -6 \\ \hline \end{array}$	$\begin{array}{r} 18 \\ -9 \\ \hline \end{array}$	$\begin{array}{r} 15 \\ -7 \\ \hline \end{array}$
Level 8	$\begin{array}{r} 15 \\ -13 \\ \hline \end{array}$	$\begin{array}{r} 19 \\ -16 \\ \hline \end{array}$	$\begin{array}{r} 14 \\ -11 \\ \hline \end{array}$
Level 9	$\begin{array}{r} 346 \\ -215 \\ \hline \end{array}$	$\begin{array}{r} 836 \\ -302 \\ \hline \end{array}$	$\begin{array}{r} 666 \\ -422 \\ \hline \end{array}$
Level 10	$\begin{array}{r} 364 \\ -3 \\ \hline \end{array}$	$\begin{array}{r} 287 \\ -11 \\ \hline \end{array}$	$\begin{array}{r} 574 \\ -133 \\ \hline \end{array}$
Level 11	$\begin{array}{r} 61 \\ -2 \\ \hline \end{array}$	$\begin{array}{r} 75 \\ -9 \\ \hline \end{array}$	$\begin{array}{r} 91 \\ -8 \\ \hline \end{array}$
Level 12	$\begin{array}{r} 36 \\ -27 \\ \hline \end{array}$	$\begin{array}{r} 47 \\ -39 \\ \hline \end{array}$	$\begin{array}{r} 75 \\ -68 \\ \hline \end{array}$

Subtraction Test (continued)

Level 13	$\begin{array}{r} 37 \\ -19 \\ \hline \end{array}$	$\begin{array}{r} 48 \\ -29 \\ \hline \end{array}$	$\begin{array}{r} 93 \\ -57 \\ \hline \end{array}$
Level 14	$\begin{array}{r} 528 \\ -64 \\ \hline \end{array}$	$\begin{array}{r} 292 \\ -84 \\ \hline \end{array}$	$\begin{array}{r} 325 \\ -32 \\ \hline \end{array}$
Level 15	$\begin{array}{r} 1067 \\ -237 \\ \hline \end{array}$	$\begin{array}{r} 4498 \\ -825 \\ \hline \end{array}$	$\begin{array}{r} 9147 \\ -735 \\ \hline \end{array}$
Level 16	$\begin{array}{r} 173 \\ -89 \\ \hline \end{array}$	$\begin{array}{r} 237 \\ -189 \\ \hline \end{array}$	$\begin{array}{r} 576 \\ -398 \\ \hline \end{array}$
Level 17	$\begin{array}{r} 700 \\ -16 \\ \hline \end{array}$	$\begin{array}{r} 900 \\ -25 \\ \hline \end{array}$	$\begin{array}{r} 600 \\ -19 \\ \hline \end{array}$
Level 18	$\begin{array}{r} 9546 \\ -7325 \\ \hline \end{array}$	$\begin{array}{r} 8132 \\ -6021 \\ \hline \end{array}$	$\begin{array}{r} 9758 \\ -8543 \\ \hline \end{array}$
Level 19	$\begin{array}{r} 8535 \\ -7986 \\ \hline \end{array}$	$\begin{array}{r} 9542 \\ -8786 \\ \hline \end{array}$	$\begin{array}{r} 6543 \\ -5754 \\ \hline \end{array}$
Level 20	$\begin{array}{r} 5941 \\ -968 \\ \hline \end{array}$	$\begin{array}{r} 6805 \\ -978 \\ \hline \end{array}$	$\begin{array}{r} 9762 \\ -986 \\ \hline \end{array}$
Level 21	$\begin{array}{r} 132428 \\ -38679 \\ \hline \end{array}$	$\begin{array}{r} 823533 \\ -245835 \\ \hline \end{array}$	$\begin{array}{r} 173461 \\ -96748 \\ \hline \end{array}$
Level 22	$\begin{array}{r} 10000 \\ -8192 \\ \hline \end{array}$	$\begin{array}{r} 80030 \\ -46759 \\ \hline \end{array}$	$\begin{array}{r} 60011 \\ -8965 \\ \hline \end{array}$

BIBLIOGRAPHY

BIBLIOGRAPHY

- Anderson, G. L. "Visual-Tactual Devices and Their Efficacy: An Experiment in Grade Eight." The Arithmetic Teacher, November 1957, pp. 196-203.
- Arnold, Felix. The Measurement of Teaching Efficiency. New York: Lloyd Adams Noble, 1916.
- Aurich, Sister M. R. "A Comparative Study to Determine the Effectiveness of the Cuisenaire Method of Arithmetic Instruction with Children of the First Grade Level." Master's thesis, Catholic University of America, 1963.
- Ayres, Leonard P. "History and Present Status of Educational Measurement." The Measurement of Educational Products, in Seventeenth Yearbook of the National Society for the Study of Education, pt. 2. Bloomington, Ill.: Public School Publishing Co., 1918, p. 9.
- Bisio, Robert M. "Effect of Manipulative Materials on Understanding Operations with Fractions in Grade V." Ed.D. dissertation, University of California, Berkeley, 1970.
- Bloom, Benjamin S., ed. Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I: Cognitive Domain. New York: Longmans, Green & Co., 1956.
- Bruner, Jerome S. The Process of Education. New York: Vintage Books, 1960.
- _____, et al. Studies in Cognitive Growth. New York: John Wiley and Sons, 1966.
- Burns, Richard W. "Achievement Testing in Competency-Based Education." Educational Technology, November 1972, pp. 39-42.
- Carmody, Lenora M. "A Theoretical and Experimental Investigation into the Role of Concrete and Semi-Concrete Materials in the Teaching of Elementary School Mathematics." Ph.D. dissertation, The Ohio State University, 1970.
- Carry, L. Ray. "A Critical Assessment of Published Tests for Elementary School Mathematics." The Arithmetic Teacher 21 (1974): 14-18.

- Carver, Ronald P. "The Coleman Report: Using Inappropriately Designed Achievement Tests." American Educational Research Journal 12 (1975): 77-86.
- Cohen, Louis. "An Evaluation of a Technique to Improve Space Perception Abilities Through the Construction of Models by Students in a Course in Solid Geometry." Ph.D. dissertation, Yishwa University, 1959.
- Cohen, Martin S. "A Comparison of Effects of Laboratory and Conventional Mathematics Teaching upon Underachieving Middle School Boys." Ed.D. dissertation, Temple University, 1970.
- Coleman, James S., et al. Equality of Educational Opportunity. 2 vols. Publication of the National Center for Educational Statistics, OE 38001. Washington, D.C.: Government Printing Office, 1966.
- Cronbach, Lee J. "Course Improvement through Evaluation." Teachers College Record 64 (May 1963): 762-683.
- Crowder, A. B. "A Comparative Study of Two Methods of Teaching Arithmetic in the First Grade." Ph.D. dissertation, North Texas State University, 1965.
- Dawson, D. T., and Ruddell, A. K. "An Experimental Approach to the Division Idea." The Arithmetic Teacher 2 (February 1955): 6-9.
- De Cecco, John P. The Psychology of Learning and Instruction: Educational Psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1968.
- Dobbin, John E. "Measuring Achievement in a Changing Curriculum." Proceedings 1956 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1957, p. 103.
- Douglass, Harl R., and Spitzer, Herbert R. "The Importance of Teaching for Understanding." The Measurement of Understanding, in Forty-Fifth Yearbook of the National Society for the Study of Education, pt. 1. Chicago: University of Chicago Press, 1946, p. 24.
- Dressel, Paul L. "Information Which Should Be Provided by Test Publishers and Testing Agencies on the Validity and Use of Their Tests: Achievement Tests." Proceedings, 1949 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1950, p. 73.
- Ebeid, William P. "An Experimental Study of the Scheduled Classroom Use of Student Self-Selected Materials in Teaching Junior High School Mathematics." Ph.D. dissertation, The University of Michigan, 1964.

- Ebel, Robert L. "Obtaining and Reporting Evidence on Content Validity." Educational and Psychological Measurement 16 (Autumn 1956): 269-282.
- _____, ed. Encyclopedia of Educational Research. 4th ed. New York: Macmillan, 1973.
- Eidson, William P. "The Role of Instructional Aids in Arithmetic Education." Ph.D. dissertation, The Ohio State University, 1956.
- Ekman, L. G. "A Comparison of the Effectiveness of Different Approaches to the Teaching of Addition and Subtraction Algorithms in the Third Grade." Ph.D. dissertation, University of Minnesota, 1966.
- Ewbank, William A. "The Mathematics Laboratory: What? When? How?" The Arithmetic Teacher 18 (1971): 559-564.
- Exner, John E., Jr. The Rorschach: A Comprehensive System. New York: John Wiley and Sons, 1974.
- Fennema, Elizabeth H. "Models and Mathematics." In Teacher-Made Aids for Elementary School Mathematics. Edited by Seaton E. Smith Jr. and Carl A. Backman. Reston, Va.: The National Council of Teachers of Mathematics, Inc., 1974, pp. 17-22.
- _____. "A Study of the Relative Effectiveness of a Meaningful Concrete and a Meaningful Symbolic Model in Learning a Selected Mathematical Principle." Technical Report No. 101. Madison: Wisconsin Research and Development Center for Cognitive Learning, 1969.
- Fitzgerald, William M., and Higgins, Jon L., eds. Mathematics Laboratories: Implementation, Research, and Evaluation. Columbus, O.: Center for Sciences and Mathematics Education, 1974.
- Glaser, Robert. "Adapting the Elementary School Curriculum to Individual Performance." Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1968, pp. 3-36.
- _____. "Psychology and Instructional Technology." In Training Research and Education. Pittsburgh: University of Pittsburgh Press, 1962.
- Green, Geraldine A. "A Comparison of Two Approaches, Area and Finding a Part of, and Two Instructional Materials, Diagrams and Manipulative Aids, on Multiplication of Fractional Numbers in Grade Five." Ph.D. dissertation, The University of Michigan, 1969.

- Green, John A. Introduction to Measurement and Evaluation. New York: Dodd, Mead and Company, 1970.
- Gronlund, Norman E. Measurement and Evaluation in Teaching, 2nd ed. New York: The Macmillan Company, 1971.
- Haggerty, M. E. "Specific Uses of Measurement in the Solution School Problems." The Measurement of Educational Products, in Seventeenth Yearbook of the National Society for the Study of Education, pt. 2. Bloomington, Ill.: Public School Publishing Co., 1918, p. 25.
- Haynes, J. D. "Cuisenaire Rods and the Teaching of Multiplication to Third Grade Children." Ph.D. dissertation, Florida State University, 1963.
- Hollis, Loye Y. "A Study to Compare the Effect of Teaching First and Second Grade Mathematics by the Cuisenaire-Gattegno Method with a Traditional Method." School Science and Mathematics 65 (November 1965): 683-687.
- Holt, John. Freedom and Beyond. New York: Dell Publishing Company, 1972.
- Howard, C. F. "Three Methods of Teaching Arithmetic." California Journal of Educational Research 1 (January 1950): 25-29.
- Howard, Vivian G. "Teaching Mathematics to the Culturally Deprived and Academically Retarded Rural Child." Ph.D. dissertation, University of Virginia, 1969.
- Johnson, Donovan A., ed. Evaluation in Mathematics. Reston, Va.: National Council of Teachers of Mathematics, 1965.
- Johnson, Randall E. "The Effect of Activity Oriented Lessons on the Achievement and Attitudes of Seventh Grade Students in Mathematics." Ph.D. dissertation, University of Minnesota, 1970.
- Judd, Charles H. "A Look Forward." The Measurement of Educational Products, in Seventeenth Yearbook of the National Society for the Study of Education, pt. 2. Bloomington, Ill.: Public School Publishing Co., 1918, pp. 159-160.
- Kieren, Thomas E. "Manipulative Activity in Mathematics Learning." Journal for Research in Mathematics Education, May 1971, pp. 228-233.
- _____. "Review of Research on Activity Learning." Review of Educational Research, October 1969, pp. 509-522.

- Kerr, Donald R., Jr. in consultation with John F. Le Blanc. "Mathematics Laboratory Evaluation." In Mathematics Laboratories: Implementation, Research, and Evaluation. Edited by William M. Fitzgerald and Jon L. Higgins. Columbus, O.: ERIC, November 1974.
- Krathwohl, David R., Bloom, Benjamin S., and Mason, Bertram B. Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook II: Affective Domain. New York: David McKay Company, Inc., 1964.
- Lankford, Frances G. "What Can a Teacher Learn About a Pupil's Thinking Through Oral Interviews?" The Arithmetic Teacher 1 (January 1974): 26-32.
- Lewy, Arie. "Discrimination Among Individuals V. Discrimination Among Groups." Journal of Educational Measurement 10 (1975): 19-24.
- Lucas, J. S. "The Effect of Attribute-Block Training on Children's Development of Arithmetic." Ph.D. dissertation, University of California, Berkeley, 1966.
- Lucon, William H. "An Experiment with the Cuisenaire Method in Grade Three." American Educational Research Journal 1 (May 1964): 159-167.
- McNemar, Quinn. Psychological Statistics. New York: John Wiley and Sons, Inc., 1969.
- Mehrens, W. A., and Lehmann, Irvin J. Standardized Tests in Education. New York: Holt, Rinehart and Winston, Inc., 1969.
- Merwin, Jack C. "Historical Review of Changing Concepts of Evaluation." Educational Evaluation New Roles, New Means, in The Sixty-Eighth Yearbook of the National Society for the Study of Education, pt. 2. Edited by Ralph W. Tyler. Chicago: The University of Chicago Press, 1969, pp. 6-25.
- Monroe, Walter S. Measuring the Results of Teaching. Boston: Houghton Mifflin Co., 1918.
- Moody, William B., Abdell, Roberta, and Bausell, Barker R. "The Effect of Activity Oriented Instruction Upon Original Learning, Transfer and Retention." Journal for Research in Mathematics Education, May 1971, pp. 208-212.
- Mott, E. R. "An Experimental Study Testing the Value of Using Multisensory Experiences in the Teaching of Measurement Units on the Fifth and Sixth Grade Level." Ph.D. dissertation, Pennsylvania State University, 1959.

- Myers, Shelton S. Mathematics Tests Available in the United States. Washington, D.C.: National Council of Teachers of Mathematics, April 1959.
- Nasea, D. "Comparative Merits of a Manipulative Approach to Second-Grade Arithmetic." The Arithmetic Teacher 13 (March 1966): 221-226.
- Nickel, Anton P. "A Multi-Experience Approach to Conceptualization for the Purpose of Improvement of Verbal Problem Solving in Arithmetic." Ph.D. dissertation, University of Oregon, 1971.
- Norman, M. "Three Methods of Teaching Basic Division Facts." Ph.D. dissertation, University of Iowa, 1955.
- Nutshall, E., and Snooh, R. "Teaching Models." In Encyclopedia of Educational Research. 4th ed. Edited by Robert L. Ebel. New York: Macmillan, 1973.
- Pace, C. R., and Stern, G. G. "An Approach to the Measurement of Psychological Characteristics of College Environments." Journal of Educational Psychology 49 (1959): 269-277.
- Passy, R. A. "The Effect of Cuisenaire Materials on Reasoning and Computation." The Arithmetic Teacher 10 (November 1963): 439-440.
- Peck, Donald M., and Jencks, Stanley M. "What the Tests Don't Tell." The Arithmetic Teacher 21 (January 1974): 54-56.
- Price, R. D. "An Experimental Evaluation of the Relative Effectiveness of the Use of Certain Multi-Sensory Aids in Instruction in the Division of Fractions." Ph.D. dissertation, University of Minnesota, 1950.
- Rankin, Paul T. "Environmental Factors Contributing to Learning." Educational Diagnosis, in Thirty-Fourth Yearbook of the National Society for the Study of Education. Bloomington, Ill.: Public School Publishing Co., 1935.
- Reavis, William C. "Contributions of Research to Educational Administration." The Scientific Movement in Education, in Thirty-Seventh Yearbook of the National Society for the Study of Education, pt. 2. Bloomington, Ill.: Public School Publishing Co., 1938, p. 27.
- Reisman, Fredicka K. A Guide to the Diagnostic Teaching of Arithmetic. Columbus, O.: Charles E. Merrill Publishing Company.
- Reys, Robert E. "Considerations for Teachers Using Manipulative Materials." The Arithmetic Teacher 18 (1971): 551-558.

- Rice, Joseph M. "The Futility of the Spelling Grind." Forum 23 (April, June 1897): 163-172, 409-419.
- Ropes, George H. "Multi-Sensory Aids in the Teaching of Arithmetic to the Second Grade." Ph.D. dissertation, Teachers College, Columbia University, 1973.
- Rugg, Harold. Statistical Methods Applied in Education. Chicago: University of Chicago Press, 1917.
- Russell, Butrand. Education and the Good Life. New York: Leveright Paperbound Edition, 1926.
- Schudson, Michael S. "Organizing the 'Meritocracy': A History of the College Entrance Examination Board." Harvard Educational Review 42 (1972): 34-69.
- Schwab, Joseph J. "The Concept of Structure in the Subject Field." Paper presented at the 20th Annual Meeting of the Council on Cooperation in Teacher Education of the American Council on Education, October 1961, Washington, D.C. Chicago: University of Chicago.
- Schwartz, Frederick J. "The Impact on Learning of COLAMADA Project Materials on Low Achievers in Mathematics." Ph.D. dissertation, University of Virginia, 1971.
- Scriven, Michael. "The Methodology of Evaluation." Perspectives of Curriculum Evaluation: American Educational Research Association, Monograph Series on Curriculum Evaluation. Chicago: Rand McNally & Co., 1967, pp. 39-83.
- Seick, Dana F. "The Value of Multi-Sensory Learning Aids in the Teaching of Arithmetical Skills and Problem Solving--An Experimental Study." Ph.D. dissertation, Northwestern University, 1959.
- Shoecraft, Paul J. "The Effects of Provisions for Imagery Through Materials and Drawings on Translating Algebra Word Problems, Grades Seven and Nine." Ph.D. dissertation, The University of Michigan, 1971.
- Simpson, Ray N. Improving Teaching-Learning Process. New York: Longmans, Green & Co., 1953.
- Sinclair, Hermine. "Piaget's Theory of Development: The Main Stages." In Piagetian Cognitive-Development Research and Mathematical Education. Edited by Myron F. Roskopf. Reston, Va.: National Council of Teachers of Mathematics, 1971.

- Skinner, B. F. About Behaviorism. New York: Alfred A. Knopf, 1974.
- _____. Beyond Freedom and Dignity. New York: Alfred A. Knopf, 1971.
- _____. Science and Human Behavior. New York: The Free Press, 1965.
- _____. The Technology of Teaching. New York: Meredith Corporation, 1968.
- Sole, David. "The Use of Materials in Teaching of Arithmetic." Ph.D. dissertation, Columbia University, 1957.
- Spross, P. M. "A Study of the Effect of a Tangible and Conceptualized Presentation of Arithmetic on Achievement in the Fifth and Sixth Grades." Ph.D. dissertation, Michigan State University, 1962.
- Squire, A., and Applebee, J. "Language Education." In Encyclopedia of Educational Research. Edited by Robert L. Ebel. New York: Macmillan, 1966.
- Stanley, J. C., and Glass, G. V. Statistical Methods in Education and Psychology. Englewood Cliffs, N.J.: Prentice Hall, Inc., 1970.
- Starch, Daniel. "Standard Tests as Aids in the Classification and Promotion of Pupils." Standards and Tests for the Measurement of the Efficiency of Schools and School Systems, in Fifteenth Yearbook of the National Society for the Study of Education, pt. 2. Chicago: University of Chicago Press, 1916, p. 143.
- Suydam, Marilyn. "Evaluation in Mathematics Classrooms: From What and Why to How and Where." ERIC. Columbus, O.: Information Analysis Center for Science and Mathematics, 1974.
- _____. "Unpublished Instruments for Evaluation in Mathematics Education: An Annotated Listing." ERIC. Columbus, O.: Information Analysis Center for Science and Mathematics, 1974.
- Swart, William L. "Evaluation of Mathematics Instruction in the Elementary Classroom." The Arithmetic Teacher 21 (January 1974): 7-11.
- Taba, Hilda. Teachers' Handbook for Elementary Social Studies. Palo Alto: Addison-Wesley Publishing Company, 1967.
- Terman, Lewis M., Lyman, Grace, Ordall, George, Ordahl, Louise E., Galbraith, Neva, and Talbert, Wilford. The Stanford Revision and Extension of the Binet-Simon Scale for Measuring Intelligence. Baltimore: Warwick and York, Inc., 1917.

- Toney, Jo Anne. "The Effectiveness of Individual Manipulation of Instructional Materials as Compared to a Teacher Demonstration in Developing Understanding in Mathematics." Ph.D. dissertation, Indiana University, 1968.
- Troyer, Maurice E. Accuracy and Validity in Evaluation Are Not Enough. New York: Syracuse University Press, 1947.
- Trueblood, Cedel R. "A Comparison of Two Techniques for Using Visual-Tactual Devices to Teach Exponents and Non-Decimal Bases in Elementary School Mathematics." Ed.D. dissertation, The Pennsylvania State University, 1967.
- Ullman, Neil R. Statistics--An Applied Approach. Lexington, Mass.: Xerox College Publishing, 1972.
- Vance, James H. "The Effects of a Mathematics Laboratory in Grade 7 and 8. An Experimental Study." Ph.D. dissertation, University of Alberta, 1969.
- _____, and Kieren, Thomas E. "Laboratory Settings in Mathematics: What Does Research Say to the Teacher?" The Arithmetic Teacher, December 1971, pp. 585-589.
- Van Engen, H. "Analysis of Meaning in Arithmetic." Elementary School Journal 49 (February-March 1949): 321-329; 395-400.
- Wasylyk, E. "A Laboratory Approach to Mathematics for Low Achievers: An Experimental Study." A working paper, University of Alberta, 1970.
- Weber, Andra W. "Introducing Mathematics to First Grade Children: Manipulative vs. Paper and Pencil." Ed.D. dissertation, University of California, Berkeley, 1969.
- Wilkinson, Jack D. "A Laboratory Method to Teach Geometry in Selected Sixth Grade Mathematics Classes." Ph.D. dissertation, Iowa State University, 1970.
- _____. "Teacher-Directed Evaluation of Mathematics Laboratories." The Arithmetic Teacher 21 (1974).
- Wolf, Richard. "The Measurement of Environments." Proceedings of the 1964 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1965, pp. 93-106.
- Wynroth, Lloyd Z. "Learning Arithmetic by Playing Games." Ph.D. dissertation, Cornell University, 1970.



3 1293 10221 5906