

INFORMATION TO USERS

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

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

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.
2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.
3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in "sectioning" the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.
4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from "photographs" if essential to the understanding of the dissertation. Silver prints of "photographs" may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.
5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

University Microfilms International

300 North Zeeb Road
Ann Arbor, Michigan 48106 USA
St. John's Road, Tyler's Green
High Wycombe, Bucks, England HP10 8HR

7821919

PAUL, ALICE SUSIANA
DEVELOPMENT OF A CLASSROOM BASED PROCEDURE
FOR ASSESSING ASPECTS OF INTELLECTUAL
FUNCTIONING OF FIRST GRADE CHILDREN.

THE UNIVERSITY OF ARIZONA, ED.D., 1978

University
Microfilms
International 300 N. ZEEB ROAD, ANN ARBOR, MI 48106

© 1978

ALICE SUSIANA PAUL

ALL RIGHTS RESERVED

DEVELOPMENT OF A CLASSROOM BASED PROCEDURE
FOR ASSESSING ASPECTS OF INTELLECTUAL
FUNCTIONING OF FIRST GRADE CHILDREN

by

Alice Susiana Paul

A Dissertation Submitted to the Faculty of the
DEPARTMENT OF ELEMENTARY EDUCATION
In Partial Fulfillment of the Requirements
For the Degree of
DOCTOR OF EDUCATION
In the Graduate College
THE UNIVERSITY OF ARIZONA

1 9 7 8

Copyright 1978 Alice Susiana Paul

THE UNIVERSITY OF ARIZONA

GRADUATE COLLEGE

I hereby recommend that this dissertation prepared under my
direction by Alice Susiana Paul

entitled Development of a Classroom Based Procedure for
Assessing Aspects of Intellectual Functioning
of First Grade Children

be accepted as fulfilling the dissertation requirement for the
degree of Doctor of Education

Kurt D. Benoit
Dissertation Director

6/5/78
Date

As members of the Final Examination Committee, we certify
that we have read this dissertation and agree that it may be
presented for final defense.

Wain Block

6/30/78

Ruth Beck

6/8/78

R. U. Allen

6/8/78

Final approval and acceptance of this dissertation is contingent
on the candidate's adequate performance and defense thereof at the
final oral examination.

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: _____

Alie S. Paul

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to her dissertation director, Dr. Keith Meredith, for his patience, direction and the assistance in the preparation of the statistical data during the preparation of this study. Appreciation is also extended to the other members of the doctoral committee for their guidance and support: Dr. Milo Blecha, Advisor; Dr. Roach Van Allen, and Dr. Ruth Beeker.

The author also acknowledges the cooperation provided by the Flowing Wells and Amphitheater School Districts, the children and their parents who participated in furnishing the data for this study. In addition, acknowledgment is made of the Tucson Early Education Model for facilitating the writer's commitment to completing this research study.

Finally, my thanks and appreciation is expressed to family and friends who provided me the continued support and encouragement to complete this extended endeavor.

TABLE OF CONTENTS

	Page
LIST OF TABLES	vi
ABSTRACT	vii
1. INTRODUCTION AND REVIEW OF LITERATURE	1
Introduction	1
Review of Related Literature	3
Definition of Intelligence	3
Operationalization of Intelligence	5
Misuse and Mistrust of Tests	8
Intellectual Development	11
Implications	12
Statement of Problem	16
2. PHASE I STUDY	17
Introduction	17
Purpose	17
Subjects	18
Instrumentation	18
Procedure	22
Results	22
3. PHASE II STUDY	29
Introduction	29
Purpose	29
Subjects	31
Instrumentation	32
Intellectual Kit Assessment Technique	32
Raven's Coloured Progressive Matrices	33
McCarthy Scales of Children's Abilities	33
Metropolitan Achievement Test	34
Procedure	35
Results	36
Reliability	36
Validity	37
Secondary	42

TABLE OF CONTENTS--Continued

	Page
4. SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS	44
Summary	44
Phase I	44
Phase II	47
Discussion and Conclusions	52
Recommendations	59
APPENDIX A: DESCRIPTION OF FORM C USED IN THE PHASE II STUDY	61
APPENDIX B: SAMPLE LETTER TO THE PARENTS	66
LIST OF REFERENCES	67

LIST OF TABLES

Table	Page
1. Item-Total Pearson Product Moment Correlations and Item Means for Forms A and B	24
2. Intellectual Skills, Item Means and Item-Total Correlations for Form C	28
3. Reliability Coefficients of Internal Consistency for IKAT	38
4. Reliability Coefficients for Raven's, MSCA and MAT	38
5. Pearson Product Correlation Coefficients between IKAT and Raven's Coloured Progressive Matrices	40
6. Pearson Product Correlation Coefficient for IKAT and MSCA	40
7. Pearson Product Correlation Coefficient for IKAT and MAT (Total Reading)	41
8. Pearson Product Correlation Coefficient for IKAT and MAT (Total Math)	41
9. Means, Standard Deviations, and Difference t-Value for Pre and Post IKAT Scores	43

ABSTRACT

The purpose of this research study was to develop an assessment procedure which would indicate a child's intellectual functioning ability in an instructional setting. This procedure was based on an instructional activity which emanates from Intellectual Kits. Intellectual Kits are defined as a collection of materials or objects which share one basic common identifying attribute, e.g., every object may be a box. However, there are also non-critical attributes such as color, size, shape and texture which permit sub-groupings of the objects. This process of observing how children perceive objects in terms of relations of similarity and dissimilarity is basic to identifying the level of children's perceptual discrimination. Children's perception of similarity and dissimilarity is a necessary foundation for children to have to deal effectively with more complex forms of discrimination and classification.

The development of the Intellectual Kit Assessment Technique (IKAT) was conducted in two phases. Phase I consisted of initial item development and item revision for the IKAT. Phase II was conducted to establish reliability and validity of the IKAT procedures resulting from the Phase I study.

Subjects were randomly selected from first grade, English speaking children from middle socioeconomic backgrounds currently enrolled in two school districts. Subjects were limited to this population to avoid additional confounding variables while trying to determine the effectiveness of assessment procedures. Thirty-six randomly selected first grade students were included in the Phase I study. A second randomly selected group of 60 first graders were subjects for the second phase of this study.

The IKAT procedure was developed to provide three scores: pre instruction score, post instruction score, and change score between pre and post instruction scores. Pretest and posttest were designed to indicate the intellectual functioning of each child at a given point in time. The change score was to indicate each child's responsiveness to instructional procedures.

Parallel forms designed for pre/post assessment of the IKAT were developed and administered in the Phase I study. There was no instruction given for this segment inasmuch as one purpose was to determine parallel form reliability. Item characteristics and internal consistency were also investigated in Phase I. Results indicated that: (1) some items did not discriminate; (2) some items were too easy or too difficult for the population; and (3) each form had low internal consistency. Based on this information and time constraints the decision was made to select the best items from the parallel forms and construct a single Form C for the pre/post assessment for Phase II.

The IKAT assessment procedure using Form C was the focus of the Phase II study. Three psychometric devices were used as criterion measures for the validation of the IKAT assessment procedure: Raven's Coloured Progressive Matrices, conceptual grouping subtest from McCarthy Scales of Children's Abilities, and Primary I Form of Metropolitan Achievement Test. This study examined the relationship among constructs as measured by IKAT and other criterion measures. Results indicated that pre/post scores of IKAT most highly correlated with intelligence tests used and to a lesser degree with measures of achievement. Validity coefficients, although lower than desired, indicated that IKAT was measuring some of the same constructs as those of the outside criterion measures with which it was compared.

A secondary result of this study provided support for Intellectual Kit instructional activity. Significant change took place between the pre and posttest on the IKAT as a result of instruction. Students performed significantly better ($p < .01$) on the posttest than on the pretest.

Further study is recommended for continued refinement of IKAT before it can be a useful classroom procedure to assess a child's intellectual functioning.

CHAPTER 1

INTRODUCTION AND REVIEW OF LITERATURE

Introduction

Assessment for the purpose of pupil guidance and for evaluating instruction should be an integral component of the teaching-learning process. Assessment within the context of the classroom is a necessary element for good teaching inasmuch as knowledge of how children are processing is essential for a teacher to know how to help each child (Dyer 1971, Silberman 1970).

Assessment devices that are available for school use are many and varied. Many of these tests appear to provide reliable and valid indications of a child's functioning in specific situations. However, there are difficulties which arise from many of the tests being administered in the schools. Feedback is delayed because the tests are usually scheduled at set times during the course of the school year and analyzed by personnel outside of the classroom. For this reason the tests are not very practical for use on a daily basis. When the test information is received by the school many teachers have difficulties interpreting the data because of a general lack of training in the specialized kinds of measurements usually used. Another difficulty arises

from most of the tests being conducted in settings isolated from the classroom (Sigel 1975). The validity of generalizations from these specific situations back to the classroom is sometimes of questionable meaning (Cazden 1974, Labov 1972, Davis 1974).

One kind of testing which has emerged in recent years is criterion-referenced testing. Attempts have been made to integrate testing with instruction through the use of criterion-referenced testing. The testing is often conducted by classroom personnel in their own learning settings. The record keeping systems incorporated provide immediate information on a child's functioning ability. This process of testing integrated with instruction has done much to alleviate some of the difficulties of testing in the schools. However, the development of criterion-referenced tests has focused on the area of achievement. Assessment in other areas of child development leaves much to be desired (Carver 1974, Davis 1974, Levine 1976, Anastasi 1976).

One critical area of concern is the child's ability to organize the surrounding environment into some meaningful schema. Authors (Hunt 1961, Gordon 1966) have referred to this ability as intellectual functioning. The teacher must consider the child's ability to function intellectually in parallel with the child's ability to function in content areas for a more complete understanding of a child's behavior. Yet, procedures for assessment of intellectual functioning abilities conducted by classroom teachers in learning settings is conspicuously lacking even

though these procedures have been and are being developed in the areas of academic achievement.

Several reasons can be found for the lack of adequate procedures for assessment of intellectual functioning in classroom settings. These would include:

1. The confusion and non-agreement concerning what intellectual functioning/intelligence is (Definition of Intelligence).
2. The operationalization of intelligence, i.e., what behavior can be termed as intelligent behavior?
3. The general misuse and mistrust of intelligence testing.
4. The various stances of how intelligence is developed (Intellectual development).

This chapter presents a review of literature in which these areas of concern are discussed. Implications are derived for assessment of intellectual functioning in school settings. The chapter concludes with a statement of the problem addressed in this study.

Review of Related Literature

Definition of Intelligence

The question of what intelligence is has raised many issues and debates among those concerned with education.

Basic to this dilemma are the numerous definitions of intelligence to which there is no overall consensus. Psychologists

are still struggling with the problem of evolving a definition or theory of intelligence that will adequately incorporate what has been learned through the use of the many tests which purport to measure intelligence (McCandless and Evans 1973, Estes 1974, Samuda 1975).

Psychological theorists of varying persuasions, such as Spearman, Thorndike, Thurstone, Cattell, and Guilford have offered numerous opinions on what the nature of intelligence is and how it is developed. The search for a definition of intelligence by the psychological theorist introduced findings and arguments which constitute individual pieces in the search for the nature of intelligence (Matarazzo 1972).

One argument has centered on the question of whether intelligence is unitary or made up of more than one factor or kind of operation (McCandless and Evans 1973). Spearman, Cattell, and Guilford have pursued the multifactor theory. Charles Spearman (in Thornburg 1973) theorized that intelligence is composed mainly of a general (g) factor which he considered the core of intelligence and a large number of specific (s) factors which are included as components. In general, his two-factor theory maintains that all mental activities have a common g factor which is the most important characteristic; each intellectual ability also has its own specific factor or factors which are closely correlated with g.

Thorndike's (in Robb, Bernardoni, and Johnson 1972) search for the nature of intelligence led him to contend that intelligence is composed of a large number of very specific elements or factors. These factors were considered to be relatively minute and could appear in combinations with other minute factors to form what seemed to be clusters of general intelligence. Thorndike hypothesized that any mental activity consisted of a great number of these minute elements operating together.

Thurstone's (in Robb et al. 1972) contribution was his discovery of primary mental abilities. He tended to discount the g factor and held that the independent factors which he labeled "group factors" were crucial to the structure of intelligence. Thurstone generalized that any complex intellectual performance was based upon a mixture of these group factors.

The contributions of each of these psychologists constitutes an important element in the ongoing process of evolving a theory of intelligence. "What has been written implies there are many ways of constructing or defining intelligence; but it is not clear that one is any more right or wrong than another (McCandless and Evans 1973, p. 161)."

Operationalization of Intelligence

Historically there has been another group of practitioner-psychologists, including Binet, Terman, Wechsler, Bayley, and Ghiselli, whose efforts have been directed to operationalizing the conceptual framework of psychological theorists. The

effort contains individuals whose activities have overlapped, but the professional activity which is labeled intellectual assessment concerning the nature of definition of intelligence has become confused with psychological testing (Psychometrics), the measurement of some product of intelligence. Binet's (in Matarazzo 1972) contribution, like that of Terman, Wechsler, Bayley, and Ghiselli provided instruments or applied techniques for the assessment and appraisal of individuals by sampling of some performance. Until Binet's contribution in 1905 the term "intelligence" as understood today was unknown. According to Wolfe (in Sarason 1976, p. 588), Binet's concept of intelligence was never conceived as having a "relatively independent existence in personality, since the weight of his writings stressed the unity of functioning in each individual." "Binet took individual differences seriously and took a dim view of premature quantification. The score on his scale observed the fact that the scale was far from comprehensive, and that what it left out was of practical significance for what one planned and did for children (Sarason 1976, p. 588)."

Other psychologists' use of Binet's conceptualization of mental age and a single scale for sampling a child's current intellectual behavior (functioning) added to the already clouded issues of the nature of intelligence. Binet's approach was seen as a global assessment of each person's intellectual capacity. Although Binet and Terman, who translated Binet's scale, saw the

test as an estimate of ability, not as a final statement, the Intelligence Quotient (IQ) has been widely interpreted as a measure of innate ability. However it is neither the only nor a complete measure of intelligence. IQ does not measure all aspects of intelligence (Matarazzo 1972). In his article "Intelligence Contra IQ" Fischer (1973, p. 12) stated the following in regard to intelligence and the use of the IQ scores:

That nineteenth century laboratory interest in psychophysical relations evolved into an elaborate quantitative methodology that produced psychology's content. Scientific psychology has failed to take human phenomena in their own right first, and then attempts to develop methods appropriate to them. Our current scientific literature, then, is about IQ, not intelligence. IQ is an artifact of the psychometric movement. It is also intimately related to Western public school criteria of educational success, which were explicitly built into the first intelligence test at the turn of the century by Binet. Specifically, IQ-test items are aimed at exposure to (a) predetermined and single perspective truth, (b) memorization of facts-as-facts, and (c) application of presented facts to logical problem-solving. In actuality, IQ-tests test for level of school achievement. IQ is then regarded as what underlies that achievement. IQ is seen as the core of intelligent being, as necessary for, and determinative of, success. This circular reduction not only keeps the low scorer down, it is also oppressive to society in general. Specifically, it gives scientific recognition and approval to only one kind of intelligence; the one-truth/analytic perspective. Competence in other approaches equally necessary for the survival of high quality society are not adequately recognized, or promoted.

Since the controversy about man's intellectual abilities and their measurement has grown and as long as the definition of intelligence remains a matter of choice, the likelihood of finding agreement on the proposition about measurement of intelligence is less likely (McCandless and Evans 1973, Samuda 1975).

Numerous group and individual tests have been developed to provide estimates of intelligence, but from the practitioner's point of view these instruments continue to cloud the nature of intelligence. Classroom teachers who are given the responsibility of making decisions about children on a day-to-day basis can find themselves being pulled between theory and practice. Much of the testing to which children are subjected in a classroom is divorced from the classroom activities and this too clouds the issues on decision-making about classroom children.

Misuse and Mistrust of Tests

In its beginning the use of educational and psychological measurement experienced a wide-spread acceptance and rather unrestricted use. More recently, however, literature and society have begun to reflect a concern about the shortcomings of measurement, such as the limitation of testing, misuses of test data, assessment techniques and even the competence and ethical responsibility of the persons who administer the tests. A major concern within the limitation of testing is that intelligence tests measure only a limited aspect of intelligence. Also such factors as socioeconomic class and educational background tend to affect test performance but are difficult to control and are often unaccounted for in determining results. Another issue relates to the fact that most tests have been developed in context of middle-class Anglo culture and are dependent upon definitions of intelligence from that culture. Further limitation can be

related to language difficulties which can invalidate the test. Misuses of test data is another category of concern particularly if test data are used to label people or to discriminate against them. Also to disregard the imprecision that is inherent in the measure and to accept a child's numerical score as a kind of absolute validity can be detrimental to the child (Silberman 1970, Robb et al. 1972, Samuda 1975, Levine 1976). The majority of achievement tests are expensive, time consuming and require trained personnel to administer them.

In part, the changes in attitude toward psychometric technology can be traced to the changes in the character of society. Psychometric concepts were developed during a time in our history when there was a great need for utilizing natural and human resources and to direct these resources in productive channels. In this context the intelligence tests provided a useful selection tool. Although American society was striving to blend varying cultural heritages into the mainstream, the intellectual competencies defined were only those held important by middle-class society (Henderson and Bergan 1976). A major portion of the individuals involved at that time were immigrants who wanted to become part of this middle-class society (Silberman 1970).

The shift today is toward a more pluralistic society which is striving to accept cultural diversity, but the widespread use of tests for purposes of selection has continued.

Donald Bersoff (1973, p. 892), in his article "Silk Purses into Sow's Ears," stated the testing situation quite aptly:

For almost 50 years, beginning with World War I, psychological testing was perceived as the vehicle by which major decisions about people's lives could be made in industry, the military, hospitals, mental health clinics, and the schools. Scores derived from psychometric instruments were used to classify, segregate, track, advance, employ, institutionalize, and educate people.

This practice has become indigenous to the kind of competitive society that characterizes all of our social institutions including educational institutions. The schools are the primary social institutions and determine the kind and amount of education for all children. This in turn determines to what extent a student will participate in the mainstream of society (Mercer 1972, McClelland 1973).

The major portion of public and private school systems do have regular testing programs. The use of tests have increased and their influence on decisions about children has become more potent and pervasive (Beggs and Lewis 1975). To alter the widespread use of tests from purposes of deciding from kindergarten on up who will fail and who will pass or who will be losers and who will be winners has not been an easy process. Since testing is part of the system, then schools should foster an attitude which will include a function of testing that will give guidance and feedback to both pupil and teacher which will improve the status of the pupil and instruction (Dyer 1971, Meyers, Ball, and Crutchfield 1973).

Intellectual Development

One of the basic issues determining the conceptual framework for the measurement of intelligence has been the question of how one views intellectual development. Initially the assumption of fixed intelligence and predetermined development dominated psychological investigation and theorizing. Since World War II there has been much debate and investigation which now supports a broader view that includes the crucial role of life experience in the development of intelligence--a view that accepts intelligence as being modifiable through environmental experience. These two viewpoints also add to the continuing argument of nature versus nurture which is not new, and there is much literature available assessing the proportional importance of each point of view (Hunt 1961, Bruner 1963, Stinchcombe 1969, Fehr 1969, Jensen 1969, Baratz and Baratz 1970, Bruner and Anglin 1973, Horowitz and Paden 1973). The implications of each of these views influences the practices related to the development, education and welfare of children and in a classroom situation each view suggests quite different approaches.

If it is assumed that intelligence is unchanging and that the course and limits of development are essentially fixed, then the task of the school and the teacher is clear; the teacher goes through the motions of teaching the child. Facts are taught as though they are known and fixed. Life is relatively simple for

the teacher because the child is blamed if progress does not occur.

If, however, development is seen as open-ended and intelligence as modifiable, then the task must become what Ira Gordon (1966) called a transactional view. From this point of view, a child is seen as highly influenced by his experience and as a contributor to society.

The latter transactional point of view also reflects the kinds of assumptions a teacher with this position would have in regard to the kind of individual the teacher would want each child to become. No one can foresee the future, but with the advances in technology there is a demand for a more responsible and problem-solving individual. It would seem that one interested in this more open point of view would opt for providing opportunities or experiences for young children to strengthen whatever potentials they have for learning. An adult would also have to provide an environment in which each child could practice becoming the kind of adult who would be productive, self-sufficient and responsible to himself and others (Stodolsky 1975, Patton 1975).

Implications

For those who accept the transactional view of the child it is imperative that the teacher has ways to assess, for practical purposes, intellectual progress measured in the classroom on a day-to-day basis. Such assessment should be more relevant for pupil guidance and for evaluating instruction. Using assessment:

information for instructional programming for individual children must become a more common practice (Hein 1975, Meier 1975). Educators have urged that we move from only using tests for predicting performance to the task of teaching to the child's needs and measuring competency change. Assisting the teacher to develop assessments that will have a direct relationship to instructional strategies seems a viable effort (Murphy 1975, Carini 1975). Therefore, to assist the practitioner involved in the day-to-day development of young children in particular, there is need for assessment tools that will assess the child's progress in order to guide and influence his development. "It is in the classroom that the content must be so arranged, and the social situation so designed, that the learner will engage himself in the operation necessary to increase his competence by mastering the content (Gordon 1966, p. 7)."

Given the transactional process as a frame of reference, adjustment in dealing with traditional assessment must be made. Traditional assessments are typically based on defined behavioral objectives which may or may not be congruent with classroom instruction. Bussi and Chittenden (1975) suggested that teaching be evaluated in terms of educational psychological constructs and not in terms of out-of-context behavioral criteria. More specifically, they (1975, p. 10) stated:

The difference between stating objectives for children and assumptions about children's capabilities and resources may seem minor at first, but it has far-reaching

implications. It is a difference that leads to: (a) a concern with environments rather than treatments; (b) an emphasis on response variability among teachers rather than response uniformity; and (c) a focus on standards of quality in learning rather than behavioral criteria outside the context of purposeful action. If research is to accommodate these priorities now being held by many educators, an overhauling of our basic paradigm seems called for.

These psychological constructs referred to by Bussi and Chittenden must be defined in such a way as to be accepted and useful in the classroom. One of the underlying expectations assumed by the numerous intelligence tests available is the demand for a child (person) to be able to respond in terms of relational meanings. It would seem practical for teachers to assess continuously the functional level of her children and their grasp of relationship (Hughes 1967). As additional support, Saunders (1973, p. 19) in his description of what intelligence test instruments measure stated the following:

If we look at the various meanings of intelligence we find that in each case the criteria have to do with: vo-
cabulary or knowing how to define many words; similarities
or knowing how to see that which is common to ostensibly
diverse things; spatial relations or being able to visu-
alize the physical relationships between things; compara-
tive analysis or the skill of seeing something as having
more meaning than meets the eye.

Assuming that a teacher is initiating varied opportunities for children which provide experiences for building understanding of relationships such as similarities and differences, spatial relations and classification then classroom assessment procedures are needed based on these same underlying constructs.

In summary, classroom teachers accept assessment as a part of instruction, but they question the appropriateness of contexts used in previous tests. Assessment should be reality based. The categories are realistic, but a change in format and the development of new procedures which can be executed by the teacher would be more appropriate for classroom use. Teachers on the whole have had to devise their own subject matter tests for student feedback. However, most of these tests, administered in the classroom, are executed as pencil and paper tasks and are content oriented. The conditions of the testing situation, whether the tests given in these situations be for achievement or intelligence, can cause anxiety or confusion on the part of the student. Labov (1972) and Cazden (1972, 1974) both reported that the situation and conditions surrounding test participation can affect the kind of feedback received from students in a testing situation. The practicality of utilizing familiar surroundings and a more informal setting with a person known to the child might result in more reliable and valid information. In addition, the use of objects or items that children have had experiences with should add relevancy to obtaining relationship information. Instrumentation is needed that will be practical and related to the classroom instructional setting. There is a need for procedures that can be incorporated into the classroom activities. Procedures conducted by the classroom teacher might indicate the students' intellectual status and also reveal their

potential for change through instruction (Hill 1963, Aldrich 1975, Bussi and Chittenden 1975, Levine 1976).

Statement of Problem

The purpose of this study was to develop a classroom based procedure which would provide teachers with an indication of a young child's intellectual functioning level in the classroom setting. This development was conducted in two phases. Phase I consisted of initial item development and item revision for the Intellectual Kit Assessment Technique (IKAT). Phase II was conducted to establish the reliability and validity of the IKAT resulting from the Phase I Study. The emphasis of this study was on the IKAT procedure itself and is only the first step needed in development before the IKAT can be adequately used as an assessment instrument in the classroom.

CHAPTER 2

PHASE I STUDY

Introduction

The purpose of this study was to develop a classroom based procedure which would provide teachers with an indication of a child's intellectual functioning ability in the classroom setting. This development was conducted in two phases. Phase I consisted of initial item development and item revision for the Intellectual Kit Assessment Technique (IKAT). The methodology and results of the Phase I study are presented in this chapter. Phase II was conducted to establish the reliability and validity of the IKAT resulting from the Phase I study. The methodology and results of the Phase II study will be presented in Chapter 3.

Purpose

Phase I was concerned with the initial development and revision of items for IKAT. Items were developed for two parallel forms, A and B. The parallel forms were to be used for pre and post assessments within IKAT. The following questions were investigated within Phase I.

1. To what degree are items within Form A and Form B of the IKAT discriminating among subjects? (Item discrimination)

2. To what extent are the items within Form A and Form B of the IKAT measuring the same construct? (Internal consistency)
3. To what degree will Form A of the IKAT correlate with Form B? (Parallel form reliability)

Subjects

Thirty-six first grade students from two school districts, Amphitheater Public Schools and Flowing Wells Public Schools, participated in this study. Twelve subjects were randomly selected from the first grade populations of each of four schools, two schools in Amphitheater and two schools in Flowing Wells school districts. Letters were sent to parents for permission to include their children in the study. The first nine letters received at each school determined the nine children in the study. Subjects in the Phase I study were Anglo from middle socioeconomic backgrounds. The subjects were limited to this population to avoid additional confounding variables while trying to determine the effectiveness of the assessment instrument.

Instrumentation

The development of the assessment forms, A and B, was based on an instructional activity which emanates from Intellectual Kits. Intellectual Kits can be defined as a collection of materials or objects which share one common identifying attribute, e.g., every object is a button. There are also many

non-critical attributes such as size, color, shape, texture or function which are not essential to the definition of the encompassing category exemplified by the set of materials. These attributes of the items readily permit subgroupings of the objects or materials (Paul, Smith, and Henderson 1970).

Through the use of Intellectual Kit material in the classroom there are recurring opportunities for children to discriminate similarities and differences between and among the objects along a number of sensory dimensions. The teacher assists the children to utilize and identify the discrimination skills used in this process. As the teacher varies the kits used in the classroom then there is opportunity for the children to generalize these discrimination skills to a variety of kits. This ability to discriminate is a necessary skill for children as they begin to order, compare, contrast and classify. This simpler level of perception however, is a necessary foundation in order for children to deal with more complex forms of discrimination. Inhelder and Piaget (1964, p. 5) as they discussed perceptual factors stated, "Long before they learn to classify objects or to arrange them in order, children perceive objects in terms of relations of similarity and dissimilarity." The discrimination process is a basis for classification and some kind of classification is implicit in a great many actions and judgments, i.e., in seriation children learn to order objects along a dimension when lining up according to size.

The proposed Intellectual Kit Assessment Technique (IKAT) incorporates the intellectual kit activity into an assessment procedure. Within the IKAT each child received a pre assessment of intellectual functioning abilities. This pre assessment was followed by a structured instructional setting focused on an intellectual kit. A post assessment was then administered to each child. The preliminary version of the IKAT contained parallel forms A and B for the pre and post assessments. Phase I was concerned with the psychometric properties of the pre and post assessments without the intervening instructional setting.

Both the pre and post assessments were based on an intellectual kit using buttons as the common object. Each assessment was composed of 12 items. The first two items were practice items to teach the format of the test. A child had to answer one of the two practice items correctly in order to be tested further.

An item consisted of a task card stimulus which was a photograph of a group of buttons. The buttons were grouped on the basis of an identifiable relationship. Each child was asked to select a button from several options which would be appropriate within that identifiable relationship. Five concrete options were available for each task for the child to choose from. The options related to the stimulus in terms of color, size, shape, texture and/or number of holes.

Scoring was dependent upon the number of characteristics in common between the child's selected response and the most

appropriate response choice. The most appropriate response choice was that choice which had the most characteristics which fit the pattern in the task. One point was given for each characteristic of the most appropriate response choice (i.e., if a child chose a button the same color as the most appropriate response choice he received one point. If he chose a button that was the same color and size as the most appropriate response choice he received two points. If he chose a button the same color, size and shape he received three points, etc.).

Each child was provided a carrel with a flannel base to use as a setting while involved with the tasks. The materials for each task included the task card, the five concrete options and an envelope into which the child placed his selected response on completion of each task.

This was not a timed test, but attention was given to item pacing to keep the children involved. Enough time was allowed for each child to choose his response. The average time for each form was 15 minutes. The subjects were tested in groups of three.

IKAT training was provided for two experienced examiners by the researcher. Four training sessions were conducted. Two of these sessions were practice sessions with children selected from the same population as those used in the study. However, these children were not included in the samples used in the Phase I and Phase II studies.

Procedure

The administration of the two parallel forms was conducted by two trained examiners in two sittings. The time between forms was held constant for all administrations. Order effects were controlled by administration of Form A then B to half of the subjects from each school district while administering Form B then A to the remaining half. Each form took approximately 15 minutes to administer. The children were tested in groups of three. Each child was provided a carrel which contained 12 tasks, each of which consisted of the photograph and five options available for response to that task. All directions were verbally presented by the examiner to each group of children. Each child's response was recorded by the examiner on the completion of each task. Since one purpose of the Phase I study was to determine parallel form reliability there was no instruction given between the administration of these two forms.

Results

Upon completion of the collection of the Phase I study data an investigation to determine alternate form reliability and internal consistency was conducted. The following statistical procedures were applied to determine if adequate reliability existed between Forms A and B as parallel forms.

The alternate form method of establishing reliability was used. This method involves the administration of two different but equivalent forms of a test to the same group of pupils. A

correlation coefficient is calculated between the two forms of the test. This coefficient is an index of the degree of agreement between pairs of standard scores for each person which is the basis for estimating the test reliability (Ebel 1969).

The correlation between Form A and Form B was .49. The means for Form A and Form B were 35.3 and 35.2. This correlation coefficient indicated that the reliability of the alternate forms was inadequate for research purposes. Related explanations for low parallel forms reliability are:

1. Each form had low internal consistency.
2. Items were too easy or too difficult for the population.
3. Items did not discriminate among the subjects.

Therefore, there was need to investigate the internal consistency and to conduct an item analysis.

Cronbach's Coefficient Alpha Index was used to determine the reliability based on internal consistency. The coefficient alpha for the two parallel Forms A and B were relatively low. Mehrens and Lehmann (1969) reported that .65 is generally acceptable for making decisions about groups. However when there is little other information on which to base a decision, it may be more desirable to use a test with low reliability than none at all.

Item analysis was then conducted to determine if specific items could be revised or eliminated in order to improve the internal consistency of the assessment instrument. Table 1 reports

Table 1. Item-Total Pearson Product Moment Correlations and Item Means for Forms A and B

Item No.	Potential Score Range	Means		Item-Total	
		A	B	A	B
1	0-5	4.66	4.58	.082	.167
2	0-5	3.97	3.50	-.058	.096
3	0-5	4.05	4.52	-.035	-.060
4	0-5	3.77	3.86	.193	.263
5	0-5	2.08	1.94	.094	-.015
6	0-5	2.16	5.0	.249	.00
7	0-5	3.97	4.16	.325	.015
8	0-5	3.00	1.86	-.168	.344
9	0-5	3.88	3.61	.074	.395
10	0-5	3.63	3.11	-.064	.175

the item-total Pearson Product Moment (PPM) Correlations and item means and the range of scores possible for each item. The PPM is an index of an item's discrimination ability. The PPM coefficients would approach 1.00 if the item is a perfect discriminator. Positive coefficients are a minimum requirement. Item means are a measure of difficulty. Ideally one would want these means to be at the middle or lower end of the potential score range prior to instruction so that growth or change, if it occurs, can be demonstrated.

Several items appeared not to be discriminating (e.g., item 3 with negative coefficients $-.035$ and $-.060$). This lack of discrimination could be due either to invalid measures of the construct, or the item's difficulty level was inappropriate for the group which limits discrimination ability (e.g., item 1 with means of 4.66 and 4.58). Either situation indicates a need for item revision or elimination.

Based on this information two alternatives were available. One alternative was to maintain the concept of parallel forms and construct new items for weak ones in each of the parallel forms. The second alternative was to select the best items from Forms A and B to construct a new common form. There were risks involved in both alternatives. To maintain the parallel form concept and construct new items would not only have involved additional cost and time, but another Phase I testing to determine the adequacy of the new items. To pursue the second

alternative and select the better discriminating items allowed estimation of its reliability from the data gathered in the first Phase I test. Since one of the purposes of this study was to investigate IKAT as an assessment instrument to indicate a child's intellectual functioning level the second alternative could still be utilized within the construct of the change score and be representative of gain from instruction when used as a pre/instruction/post unit.

Due to limited time and expense constraints the decision was made not to construct new items but to select the better discriminating items from the parallel forms in order to obtain a single form to be called Form C. A total of 12 items, two practice items and the 10 principal items were selected for the new form.

The estimated internal consistency of the new Form C using Phase I data was .62. Although lower than desired, it was higher than Forms A and B previously.

The criteria used for selection of the items were:

(1) the item had to have a mean less than 4.00, and (2) the item had to have a discrimination index of .20 or greater. Two exceptions were made in the selection. One item (Form C, item 7) was selected because the researcher wanted the subjects to have some success during the assessment. This item, when reanalyzed with the selected items, had a marginally acceptable discrimination index (.199). The other item (Form C, item 5) was selected

because when this item was included in the reanalysis of the data the internal consistency was higher than without the item.

Therefore, the item was included even though the discrimination index (.176) was below the .20 cutoff. Table 2 reports the intellectual skill, the item mean and item-total correlation for each of the items selected for Form C. The item-total correlations were based on the assumption that if these items were combined the student responses would remain the same. In this regard they must be considered estimates.

The Phase II study was conducted using Form C for both pre and post assessment based on the results of the Phase I study.

Table 2. Intellectual Skills, Item Means and Item-Total Correlation for Form C

Item No.	Intellectual Skills	Means	Item-Total
1	Discrimination (color)	3.77	.317
2	Discrimination (pattern relationship)	2.08	.329
3	Discrimination (paring)	2.16	.397
4	Discrimination (texture)	3.97	.202
5	Discrimination (seriation)	3.50	.176
6	Discrimination (seriation)	3.86	.340
7	Generalization	4.16	.199
8	Inference	1.86	.349
9	Inference	3.61	.368
10	Inference	3.11	.380

CHAPTER 3

PHASE II STUDY

Introduction

The methodology and results of the Phase II study regarding the investigation of reliability and validity are reported in this chapter. The primary purpose of the Phase II study was to investigate the validity of the IKAT assessment procedure. In addition, reliability coefficients were reestablished for the IKAT pre, post, and change scores as only estimates were obtained from the Phase I study. No reliability coefficients were obtained for the IKAT change score. Reliability coefficients were also established for the Raven's and the McCarthy's Scales.

Purpose

Reliability is essential before validity can be established for any measure. Reliability can be determined in a number of ways: internal consistency, test-retest and parallel form reliability. Internal consistency was determined to be the most appropriate reliability type for the design of this study for two reasons: parallel forms are not being used and the reliability of change across time on the IKAT is being assessed through the

IKAT change score. The following three questions regarding internal consistency were investigated and will be addressed in turn:

1. To what degree are the items within Form C of IKAT measuring a common construct? This reliability was estimated for Form C in the Phase I study. However, this was only an estimate and must be reestablished in the Phase II study.
2. To what degree are the items within the Raven's measuring a common construct? The Cronbach Alpha reliability coefficient was computed for the Raven's as no coefficient was reported in the Raven's Manual or in Buros' Sixth Mental Measurement Yearbook (Bortner 1965).
3. To what degree are the items within the conceptual grouping subtest of MSCA measuring a common construct? Internal consistency coefficients were reported for the Perceptual Performance Scale of the MSCA but not for the conceptual grouping subtest which was the subtest used in this study. Therefore the Cronbach Alpha reliability coefficient was computed for the conceptual grouping subtest of MSCA.

Internal consistency coefficients for the MAT subtests were not computed as there are extensive data from several studies establishing the reliability coefficients of the MAT subtests across situations and populations.

Validity is the ultimate question which must be investigated for any new assessment technique. Validity can be investigated several ways. The most common procedure is to investigate the interrelationships between the proposed assessment and other criterion measures. The following criterion-related validity questions were investigated in the Phase II study:

1. To what degree will the IKAT scores correlate with the conceptual grouping subtest of the McCarthy Scales of Children's Abilities?
2. To what degree do IKAT scores correlate with the Raven's Coloured Progressive Matrices?
3. To what degree do IKAT scores correlate with the Metropolitan Achievement Test?

Subjects

Sixty first grade students from two school districts, Amphitheater Public Schools and Flowing Wells Public Schools, participated in this study. Fifteen subjects were randomly selected from the first grade population of each of four schools, two schools in Amphitheater and two schools in Flowing Wells school districts. Those students involved in the Phase I study were not included in this population. Letters were sent to parents for permission to include their children in the study. The subjects were Anglo from middle socioeconomic backgrounds as in the Phase I study.

Instrumentation

The IKAT assessment procedure using Form C developed in the Phase I study was the focus of this validation study. Three psychometric devices were used to validate this assessment procedure, the Raven's Coloured Progressive Matrices, the conceptual grouping subtest from the McCarthy Scales of Children's Abilities, and the Primary I Form of the Metropolitan Achievement Test.

Intellectual Kit Assessment Technique

The IKAT consisted of administering Form C to three children at a time as a pretest. The group of children was then introduced to an instructional session.

The instructional session was structured to include the classification skills assessed by Form C but were generalized to another kit. The child was directed through a series of questions and tasks or a combination of tasks that would take him through a process of classification using a spoon kit. Form C was based on a button kit. Following this instructional session the group was then administered Form C as a posttest. The items, administration procedures, and scoring procedures for Form C were described in Chapter 2. A copy of Form C is found in Appendix A.

The IKAT provided three scores: the pre-instruction score, the post-instruction score, and the change score between pre and post instruction scores. The pretest and the posttest were to indicate the intellectual functioning of each child at a point in time. The change score was to indicate each child's

responsiveness to instruction. Reliability and validity of each of these scores was investigated.

Raven's Coloured Progressive Matrices

The Raven's Coloured Progressive Matrices (RCPM) developed by Raven (in Buros 1965) were designed to assess the main cognitive processes of which children under age 11 are usually capable.

These tests represent an attempt to measure intellectual functioning within the context of Spearman's concept of "g." The tasks or matrices consists of designs which require completion. The testee chooses from multiple choice options the design or design part which best fits. An answer which fits may: (a) complete a pattern, (b) complete an analogy, (c) systematically alter a pattern, (d) introduce systematic permutations, or (e) systematically resolve figures into parts (Bortner 1965, p. 491).

The review in Buros' Sixth Mental Measurement Yearbook by Bortner (1965, p. 490) makes only one reference to reliability and validity: "The accumulating literature dealing with the validity and reliability of these scales is equivocal." However, it is a commonly used test with young children. Its general use and availability for assessment of cognitive processes provides a source of comparison and validation for this study.

McCarthy Scales of Children's Abilities

The McCarthy Scales of Children's Abilities (MSCA) were designed to meet the need expressed by psychologists for a single instrument to assist in evaluation of the young child's general

intellectual level of functioning and specific patterns of strength and weaknesses in important abilities.

Measurement by the MSCA is suitable for children between the ages of two and one-half and eight and one-half years. The McCarthy Scales consist of 18 separate tests which have been grouped into six scales: Verbal, Perceptual-Performance, Quantitative, General Cognitive, Memory, and Motor. The subtest test to be used in this study is the conceptual grouping subtest of the Perceptual-Performance Scale. A child is required to classify blocks on the basis of size, color, and shape. The child's ability to deal logically with objects, to classify and to generalize is assessed through his manipulation (McCarthy 1972). There is no reliability coefficient reported for the conceptual grouping subtest alone, but only as part of the Perceptual-Performance Scale; however, internal consistency will be examined in this study. This subject was chosen as a comparison for validity as the tasks are very similar to the tasks to be used in this study.

Metropolitan Achievement Test

Metropolitan Achievement Test (MAT) is designed to evaluate what is being taught in today's schools. The development of content for the tests is based on extensive analysis of current curricular materials such as leading testbook series, syllabuses and state guidelines.

The MAT is standardized on a national basis. The primer was standardized twice during one year in January and April of 1970. Subjects tested in kindergarten were selected from a subsample of schools for which the median IQ in grade one was 100. The standardization samples were selected to represent the national population in terms of geographic region, size of city, socioeconomic status, and public versus non-public schools.

Reliability data were reported in split-half (odd-even) coefficients, corrected by the Spearman-Brown formula. For the end of kindergarten the reliability coefficients range from 90 to 92 for the norm group. For the middle of grade one the reliability coefficients range is from .89 to .93 (Durost et al. 1971).

Procedure

During the first setting each subject was administered a pretest with Form C of IKAT by a trained examiner. Upon completion of the pretest the subjects were engaged in an instructional activity using a Spoon Intellectual Kit by the researcher. Following this instructional activity an examiner then administered a posttest to each subject using Form C. The total administration time for this technique was approximately 40 minutes. Children were tested in groups of three.

At a second setting each child was administered the conceptual grouping subtest of the McCarthy Scales of Children's

Abilities (CG) followed by Raven's Coloured Progressive Matrices (Raven) by an examiner. The total administration time for this portion was 30 minutes.

The Metropolitan Achievement Test (MAT) was administered two weeks later in three one-half hour settings. All children were administered the MAT but in groups of less than 10.

Results

Reliability

In the Phase I study the internal consistency coefficients for the two parallel Forms A and B were relatively low. Mehrens and Lehmann (1969) reported that it is generally accepted that tests used to assist in making decisions about individuals should have reliability coefficients of at least .85. Reliability coefficients of about .65 are acceptable for making decisions about groups. However another factor to be considered is how good a decision can be made without the help of any test data. "If there is very little other information on which to base a decision, and a decision must be made, it may be helpful to use a test with low reliability rather than none at all. A test with low reliability can still have some validity and can therefore be useful (Mehrens and Lehmann 1969, p. 41)." The decision was made to select the better discriminating items from both Forms A and B for a single form to be called C. The items selected for Form C had both variability and a higher

correlation with the total score. This procedure allowed the estimation of Form C's reliability from the data gathered in the Phase I study. The estimated internal consistency of the new Form C using Phase I data was .62. Given that this was only an estimate of internal consistency the need to reestablish the reliability for Form C was necessary.

Table 3 reports the Cronbach Alpha coefficients of internal consistency for the IKAT Pretest score, Posttest score and the Gain score. The revision and selection of items from Forms A and B did result in a higher internal consistency for Form C.

Table 4 reports the computed reliability coefficients for the Raven and MSCA conceptual grouping subtest. The MAT reliability coefficients reported are from the MAT test manual.

The question of whether these coefficients are higher or lower is unanswerable. Coefficients are relative to other instruments measuring similar constructs. For this study the question is not how high or low but rather what limitations will these reliability coefficients place on the validity coefficients yet to be established.

Validity

Pearson Product Moment Correlation coefficients (PMA) were computed for each combination of IKAT score and criterion measure. The correlation coefficients among IKAT scores, Raven's,

Table 3. Reliability Coefficients of Internal Consistency for IKAT

	Pre	Post	Gain
Reliability Coefficients	.54	.65	.56

Table 4. Reliability Coefficients for Raven's, MSCA and MAT

	Raven's	MSCA	MAT
Reliability Coefficients	.80	.54	.96 Total Reading
			.93 Total Math

MSCA, MAT and the correlation coefficients corrected for attenuation are reported.

In examining validity coefficients if we had perfectly reliable instruments (Cronbach Alpha = 1.00) then validity coefficients may also approach 1.00; however to the degree that reliability coefficients are less than 1.00, validity coefficients will also be limited. For example, if our reliability coefficient was .65, then the upper limit of our validity coefficient would be .81 (Cronbach 1949). Limitations must be taken into account in interpreting validity coefficients. One would like to know what the relationship among the constructs is without the reliability limitation.

The correction for attenuation was used to estimate what the correlation would be if variables were perfectly reliable (Nunnally 1967). This was an appropriate procedure since the interest of this study was to examine the relationships among constructs as measured by IKAT and the other criterion measures.

Table 5 reports the Pearson Product Correlation coefficients uncorrected and corrected for attenuation between IKAT scores and the Raven's. Table 6 reports the Pearson Product Correlation Coefficient uncorrected and corrected for attenuation between each IKAT score and MSCA.

Tables 7 and 8 report the Pearson Product Correlation Coefficient uncorrected and corrected for attentuation between each IKAT and the MAT. Total Reading is reported in Table 7 and Total Math reported in Table 8.

Table 5. Pearson Product Correlation Coefficients between IKAT and Raven's Coloured Progressive Matrices

IKAT Scores	Correlation	Correlation Corrected for Attenuation
Pre	.41	.62
Post	.33	.46
Gain	-.06	-.09

Table 6. Pearson Product Correlation Coefficient for IKAT and MSCA

IKAT Scores	Correlation	Correlation Corrected for Attenuation
Pre	.13	.24
Post	.07	.12
Gain	-.05	-.09

Table 7. Pearson Product Correlation Coefficient for IKAT and MAT (Total Reading)

IKAT Scores	Correlation	Correlation Corrected for Attenuation
Pre	.22	.31
Post	.29	.37
Gain	.08	.11

Table 8. Pearson Product Correlation Coefficient for IKAT and MAT (Total Math)

IKAT Scores	Correlation	Correlation Corrected for Attenuation
Pre	.30	.42
Post	.35	.45
Gain	.06	.08

The summary tables 5 through 8 indicate that the pre/post scores of IKAT most highly correlate with the intelligence tests and to a lesser degree with measures of achievement. The correlation coefficients were highest with the Raven's; however, there was a low correlation with MSCA. The low relationship with the MSCA can be attributed to the restricted range of MSCA. This instrument was relatively easy for the population tested. A restricted range of scores with a measure will result in a lower correlation with other measures. Within the achievement measures the IKAT related more to the math section of the MAT than the reading section. The math section of the MAT deals with relationships similar to IKAT.

Secondary

Within many studies there are often results which were not a part of the original questions investigated, but need to be reported as new information. Such a finding exists in this study. The Spoon Intellectual Kit was used for instruction between the pre/post segment of IKAT and significant change took place on the IKAT as a result of the instructional sequence. Students performed significantly better ($p < .01$) on the posttest as compared to the pretest. Means, standard deviations, and the resulting t-value are reported in Table 9.

Table 9. Means, Standard Deviations, and Difference t-Value for Pre and Post IKAT Scores

	Pre	Post	Difference	t-Value
Mean	36.97	39.95	2.98	4.58
S.D.	4.47	4.69	5.06	

CHAPTER 4

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

This final chapter contains a summary of the research and findings for Phase I and Phase II Studies, the conclusions formulated as a result of each study and the recommendations for further research.

Summary

The overall purpose of this study was to develop a classroom based procedure which would eventually provide teachers with an indication of a child's intellectual functioning ability in the classroom setting. This development was conducted in two phases. Phase I consisted of initial item development and item revision for the Intellectual Kit Assessment Technique (IKAT). Phase II was conducted to establish the reliability and validity of the IKAT resulting from the Phase I study.

Phase I

Purpose. Phase I was concerned with the initial development and revision of items for IKAT. Items were developed for two parallel forms, A and B. The parallel forms were to be used for pre and post assessment within the IKAT. The following questions were investigated within Phase I:

1. To what degree are items within Form A and Form B of the IKAT discriminating among subjects? (Item discrimination.)
2. To what degree are the items within Form A and Form B of the IKAT measuring the same construct? (Internal consistency.)
3. To what degree will Form A of the IKAT correlate with Form B? (Parallel form reliability.)

Subjects were first grade students from two school districts in the Tucson area. The parallel assessment forms, A and B, were based on an instructional activity which emanates from Intellectual Kits. The Intellectual Kit Assessment Technique (IKAT) incorporates the Intellectual Kit activity into an assessment procedure. Phase I was concerned with the psychometric properties of the pre and post assessments without the intervening instructional settings.

The two parallel forms were administered in two settings to groups of three students each. Each form took approximately 15 minutes to administer.

Results. An investigation to determine alternate form reliability and internal consistency was conducted.

Reliability of the alternate forms were inadequate for research purposes. Related explanations for low parallel forms reliability are:

1. Each form had low internal consistency.
2. Many items were too easy or too difficult for the population.
3. Items did not discriminate among the subjects.

Therefore, there was need to investigate the internal consistency and to conduct an item analysis.

Cronbach's Coefficient Alpha Index was used to determine the reliability based on internal consistency. The Coefficient Alpha for the two parallel Forms A and B were relatively low. Mehrens and Lehmann (1969) reported that .65 is generally acceptable for making decisions about groups. However, when there is little other information on which to base a decision, it may be more desirable to use a test with low reliability than none at all.

Item analysis was then conducted to determine of specific items could be revised or eliminated in order to improve the internal consistency of the assessment instrument. Several items appeared not to be discriminating. Rather than maintain the concept of parallel forms and construct new items for weak ones in each of the forms the decision was made to select the better discriminating items from both Forms A and B for a single form to be called C. Form C's reliability was then estimated from the data gathered in the Phase I study. The estimated internal consistency of the new Form C using Phase I data was .62. Although lower than desired, it was higher than either Forms A and B previously.

This Form C was then used in the Phase II study for both pre and post assessment with instruction intervening.

Phase II

Purpose. The primary purpose of the Phase II study was to investigate the validity of the IKAT assessment procedure. In addition, reliability coefficients were established for the IKAT pre, post, and change scores. Only estimates were obtained from the Phase I study and needed to be reestablished.

Reliability is essential before validity can be established. Internal consistency was determined to be the most appropriate reliability type for the design of this study for two reasons: parallel forms were not being used and the reliability of change across time on the IKAT was being assessed through the IKAT change score. The following three questions regarding internal consistency were investigated and were addressed in turn:

1. To what degree were the items within Form C of IKAT measuring a common construct? (Internal consistency.)

This reliability was estimated for Form C in the Phase I study. However, this was only an estimate and was reestablished in the Phase II study.

2. To what degree were the items within the Raven's measuring a common construct? (Internal consistency.) The Cronbach Alpha reliability coefficient was computed for the Raven's as no coefficient was reported in the Raven's

Manual or in Buros' The Sixth Mental Measurement Yearbook (Bortner 1965).

3. To what degree were the items within the conceptual grouping subtest of MSCA measuring a common construct? (Internal consistency.) Internal consistency coefficients were reported for the Perceptual Performance Scale of the MSCA but not for the conceptual grouping subtest which was the subtest used in this study. Therefore, the Cronbach Alpha reliability coefficient was computed for the conceptual grouping subtest of MSCA.

Internal consistency coefficients for the MAT subtest were not computed as there is extensive data from several studies establishing the reliability coefficients of the MAT subtests across situation and populations.

Validity is the ultimate question which must be investigated for any new assessment technique. Validity can be investigated several ways. The most common procedure is to investigate the interrelationships between the proposed assessment and other criterion measures. The following criterion-related validity questions were investigated in the Phase II study:

1. To what degree did the IKAT scores correlate with the conceptual grouping subtest of the McCarthy Scales of Children's Abilities?

2. To what degree did IKAT scores correlate with the Raven's Coloured Progressive Matrices?
3. To what degree did IKAT scores correlate with the Metropolitan Achievement Test?

Subjects were 60 first grade students from two school districts in Tucson, Those students involved in the Phase I study were not included in this population.

The IKAT assessment procedure using Form C developed in the Phase I study was the focus of this validation study. Three psychometric devices were used to validate this assessment procedure, the Raven's Coloured Progressive Matrices, the conceptual grouping subtests from the McCarthy Scales of Children's Abilities and Primary I form of the Metropolitan Achievement Test.

During the first setting a pretest with Form C of IKAT was administered by a trained examiner to each subject. Upon completion of the pretest the subjects were engaged in an instructional activity using a Spoon Intellectual Kit by the researcher. Following this instructional activity an examiner then administered a posttest to each subject using Form C. The total administration time for this technique was approximately 40 minutes. Children were tested in groups of three.

At a second setting each child was administered the conceptual grouping subtest of the McCarthy Scales of Children's Abilities (CG) followed by the Raven's Coloured Progressive

Matrices (Raven's) by an examiner. The total administration time for this portion was 30 minutes.

The Metropolitan Achievement Test (MAT) was administered two weeks later in three one-half hour settings. Children were administered the MAT in groups less than 10.

Results--Reliability. Cronbach Alpha coefficients of internal consistency for the IKAT Pretest score, Posttest score and the Gain score were reestablished after the revision and selection of items from Forms A and B. Higher internal consistency coefficients were obtained for Form C IKAT scores than was obtained previously for either Forms A or B. Cronbach Alpha coefficients of internal consistency for the Raven's was computed as there was no reliability reported in the Raven's Manual. The reliability coefficient was reported for the Perceptual Performance Scale of the MSCA but not for the conceptual grouping subtest which was the subtest used in this study. Therefore, the Cronbach Alpha reliability coefficient was computed for the conceptual grouping subtest of MSCA (.54). The internal consistency coefficients for MAT total reading (.96) and total math (.93) as reported in the MAT test manual were used.

The question of whether these coefficients are high or low is unanswerable. Coefficients are relative to other instruments measuring similar constructs. For this study the question was not how high or low but rather what limitations would these reliability coefficients place on the validity coefficients yet

to be established. These limitations were taken into account and were adjusted for through the use of the correction for attenuation.

Pearson Product Moment Correlation coefficients (PPMC) were computed for each combination of IKAT score and criterion measure. The correlation coefficients corrected for attenuation were also computed.

The correction for attenuation was used to estimate what the correlation would be if variables were perfectly reliable (Nunnally 1967). This was an appropriate procedure since the interest of this study was to examine the relationships among constructs as measured by IKAT and the other criterion measures.

Data analyses indicate that the pre/post scores of IKAT most highly correlate with the intelligence tests and to a lesser degree with measures of achievement. The correlation coefficients were highest with the Raven's. Relatively low coefficients were obtained for interrelationships between IKAT scores and the MSCA. The low relationships with the MSCA can be attributed to the restricted range of MSCA. This instrument was relatively easy for the population tested. A restricted range of scores for a measure will result in a lower correlation with other measures. Within the achievement measures the IKAT related more to the math section of the MAT than the reading section. The math section of the MAT deals with relationships similar to those found on the IKAT.

Discussion and Conclusions

The reliability data computed for the pre/post and change scores of IKAT were lower than desired both in the Phase I study and Phase II study. Factors which may have contributed to the unreliability of all IKAT scores would include test length, group homogeneity, item selection, options provided, format of instrument, and the accuracy of the picture stimuli provided. A discussion of these factors follows.

The length of IKAT is short due to the time requirement for the entire IKAT assessment procedure (pre, instruction, post). The reliability of scores obtained on a sample of items increase with the number of items sampled. One primary way to make tests more reliable is make them longer (Nunnally 1967). As the change scores did not correlate with any other measures one could question whether or not the entire process of pre/instruction/post measurement format is needed. If the instructional and posttest segments were eliminated the IKAT form could be tripled in length and still remain in the same time frame. Using the Spearman-Brown Prophecy formula the estimated reliability for the IKAT assessment tripled in length would be .78 (Mehrens and Lehmann 1969).

Group homogeneity refers to the relationship between group variability and reliability. Reliability is restricted to the degree that group variability is limited (Mehrens and Lehmann 1969). Homogeneity would be indicated by easy or difficult items

with limited variability. Some of the selected items used in IKAT were relatively easy, but were recording knowledge of basic concepts. The easier items insured some success to the children involved and these items need to be retained.

Reliability of the IKAT could be improved through further refinement of the options. This refinement would increase the item's ability to discriminate. Chase (1974) suggested that items can be improved by making sure all options are reasonable and attractive to the item stem. Discrimination ability of items can also be improved by increasing the number of options, however, one may question whether children of this age can discriminate among the many options for IKAT. A four-option format may be better for this particular population instead of the five-option format used. In addition, different weighting (or scoring) procedures need to be further investigated.

The format of the test itself may have been an influencing factor. The use of an abstract stimulus (pictures) may have been too demanding for some of the children. The instructional phase of the IKAT process used concrete objects. The scores of the majority of children were higher as a result of the instructional phase. Therefore, the abstract-concrete format appeared to be appropriate for many of the children. An alternative method for further study could be to develop two phases of testing which would use a concrete-concrete format followed by an abstract to concrete which might provide a more natural transition

of tasks for children of this age. The first phase would use concrete items not only as options but also as the task. The second phase would then move from an abstract stimulus to concrete options which is the present format of IKAT. The photographic stimuli for each task could also have detracted from the tasks. There was considerable time and effort spent with the film developers in trying to achieve accurate representation of the buttons particularly in color and contrast. However, the desired accuracy was never obtained.

In this study as in other studies of assessment procedures there are numerous sources of error variance which could cause the performance scoring of individuals to vary. Conditions such as trait instability, sampling error, administration error, scoring error, health, motivation, degree of fatigue or even good or bad luck in guessing can account for differences in scores (Mehrens and Lehmann 1969). The fewer the errors the more reliable the measurement will be. Further refinement of standardized procedures would aid in controlling these reliability factors.

Change scores reliability and the measure of same has typically been a problem for the researcher (Linn and Slinde 1977). The unreliability of change scores are more often due to computational problems inherent in measuring change rather than to unreliability of change itself. It would seem that a natural measure of change from one point in time to another is the simple difference score between the two measurement points. However,

this procedure results in a score with several limitations that are inherent in the data.

One major defect in the measurement of the change score is that it typically has a negative correlation with the pretest. In order for the correlation to be positive the standard deviation of the post measure must be substantially larger than the pre measure and the correlation between pre and post must be less than one. Usually this is not the case as the standard deviation of the pre and post measures are usually relatively similar in magnitude and the pre-post correlation often approaches 1.00.

The reliability of the change score is dependent on two characteristics of the pre/post measures on which it is based:

1. Reliability of pre/post measures.
2. Correlations between pre/post measures.

As the reliabilities of the pre/post measures increase the reliability will increase. However, as the correlations between pre/post measures increase the reliability of the change score will be decreased. Maximum reliability of the change score will be obtained when the pre and post measures have high internal reliability and low intercorrelation. Yet, if this circumstance exists, one would question if the pre and post measures are measuring the same construct. The same constructs must be measured by the pre and post measures, or the validity of the pre, post, and change scores will be questionable. Several

procedures have been examined in attempts to eliminate these inherent problems of measuring change. After a critical review of these procedures, Richards (1975) concluded that simple difference scores are as good as more complicated methods.

Criterion-related validity was investigated through Pearson Product Moment Correlation coefficient (PPMC) among Raven's, the conceptual grouping subtest of the MSCA, the Primary I form of the MAT and the IKAT scores. Correlation coefficients corrected for attenuation were also computed for each combination of IKAT score and criterion measure. These coefficients were lower than desired in that a considerable amount of variance is unaccounted for. However, considering the age level of the population tested, it is understandable that there could be a number of sources contributing to test variability.

The correlation coefficients, uncorrected and corrected for attenuation between IKAT scores and the Raven's were moderate but were higher than the correlation coefficients of the other outside criterion measures used.

The Pearson Product Correlation coefficients between IKAT scores and MSCA were low. Only a portion of the conceptual group subtest was used. Two difficulties arose because of this selection:

1. The nine items making up this portion were too easy for a large number of subjects which resulted in a very skewed

distribution. This may have contributed to the low correlations.

2. This portion had low internal consistency because of both the level of difficulty and the limited number of items. This, too, would have limited the intercorrelations.

The correlation coefficients for both the total reading and math of the MAT and the IKAT were moderate and lower than the correlation coefficients analyzed for the Raven's. As such the higher correlation of the IKAT and the Raven's could be indicative that the IKAT is more an intellectual ability measure than an achievement measure.

Although the validity was lower than desired in both the Phase I study and the Phase II study it was indicative that IKAT was measuring some of the same constructs as the outside criterion with which it was compared. As was discussed the unexplained variance encountered could be due to difficulties in assessing abilities of young children.

The change score remains a puzzle in that it did not correlate with the other measures in this study, yet the internal consistency coefficient provides evidence that there is an underlying construct being assessed.

A secondary finding of this research study provides empirical evidence which supports the validity of the use of the Intellectual Kit as an effective instructional procedure. The Intellectual Kit instructional procedure has been utilized as a

training tool for teachers and a teaching tool for children in the Tucson Early Education Model for a number of years and support for the kit has been on an informal experiential basis.

There now exists supporting evidence which has been lacking in available literature. Two specific pieces of evidence include:

1. There was significant change in the IKAT scores as a result of the instructional sequence.
2. That the learning attained from one set of objects, e.g., Spoon Kit, can generalize to another set, e.g, Button Kit.

Activities organized and planned around the use of Intellectual Kits would appear to be justified because children change the nature of their responses after experiencing the activities of the kits.

However, two limitations exist in this study which require replication studies of the findings:

1. There was no control group as the study was primarily a reliability and validity study of the IKAT assessment procedure; and
2. The same instrument was used as the pre and post assessment instrument which may have resulted in a testing effect, i.e., the children may have learned from taking the pretest rather than the instructional sequence.

Recommendations

Upon completion of this study the following recommendations are expressed by the researcher.

Before IKAT can be considered a viable tool for assessing a child's intellectual functioning level further research is necessary. The following steps should be considered as recommendations:

1. Additional studies need to be conducted in a continued effort to improve IKAT as an assessment instrument. Improvements would include lengthening the test, refining the options, improving the scoring, changing the format, improving the accuracy of the stimuli.
2. Additional validity studies need to be conducted in order to determine the underlying construct(s) being measured by the IKAT scores including the change score.
3. Research should be conducted which will determine whether teachers can use the IKAT in actual instructional settings and make more informed decisions based on the resulting data.
4. Research of IKAT as an assessment instrument should be conducted with different children in various settings to determine its generalizability.

The continued refinement of IKAT may provide a viable instrument which the classroom teacher may use to assess

intellectual functioning. This assessment may be used to provide a focus for instruction without labeling the child.

APPENDIX A

DESCRIPTION OF FORM C USED IN THE PHASE II STUDY

Form C evolved from the intended parallel Forms A and B which were used in the Phase I study. Form C contains the items which had both variability and a higher correlation with the total score.

The children were tested in groups of three. Each child had a carrel which contained 12 envelopes with photographed tasks and five options. Each envelope was numbered from 1 to 12. All directions were verbally presented by the examiner. Each child's response was recorded by the examiner upon completion of each task.

The following is a description of the task and directions given by the examiners.

The stimuli for each item was a photograph depicting selected buttons which indicated a task relating to color, size, shape, texture and number of holes.

Practice Item A: Take out the first envelope and picture out of the pocket. Empty the envelope on the front part of your box (point to this area on the carrel). Turn the buttons and the picture right side up. Let each child

examine their buttons, then say--There are many different kinds of buttons. Are all these buttons the same? How are they different? Discuss sizes, shapes, colors, number of holes. Now say--look at the buttons in the picture. Find the button that is most like the ones in the picture. Put it beside in the picture. (Not on top.) Check each child's response, then say--That's very good, now put all the buttons back in the envelope and put the envelope and picture back in the pocket.

Practice Item B: Take out the second envelope and picture out of the pocket. Empty the envelope on the front part of your box (point to this area on the carrel). Turn the buttons and picture right side up--This time we're going to do something a little different. Look at your picture. There is an arrow pointing left to right. These buttons are in a pattern. The arrow points the direction the pattern goes. Look at the picture and when you discover the pattern find the button that goes next in line (point to the place in the photograph). From now on, when you see an arrow remember that it shows which way a pattern goes. When you discover the pattern for this picture, find the button that goes next in line.

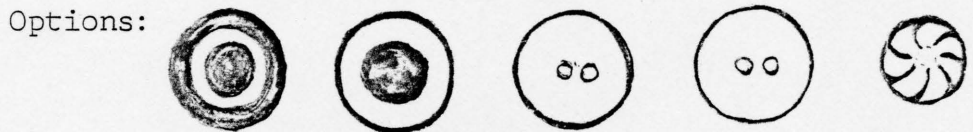
Item 1. Task photographed: Three textured cloth covered buttons of varying design and size.

Directions: This is a set. The buttons go together for some reason. Discover the reason. Find the button that goes with the set for the same reason.



Item 2. Task photographed: Three identical buttons.

Directions: Wait for me to tell you what to look for. Find the button like those in the picture but a different color.



Item 3. Task photographed: Four shank buttons same size, texture, but each a different color.

Directions: Find the button that goes with this set.



Item 4. Task photographed: The textured cloth buttons varying size/designs.

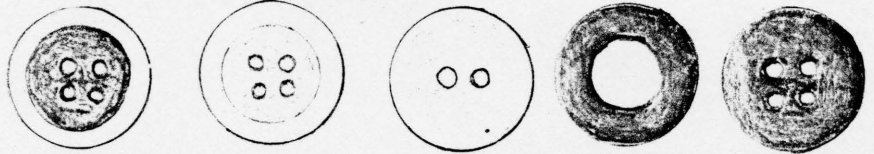
Directions: This is a set. The buttons go together for some reason. Discover the reason. Find the button that goes with set for the same reason.



Item 5. Task photographed: Three identical buttons.

Directions: Wait for me to tell you what to look for.
Find the button like those in picture but
a different color.

Options:



Item 6. Task photographed: Three buttons different size/shape/
color/number of holes/no holes.

Directions: These buttons go together for some reason.
When you discover the reason, find the
button that goes with the set for the same
reason.

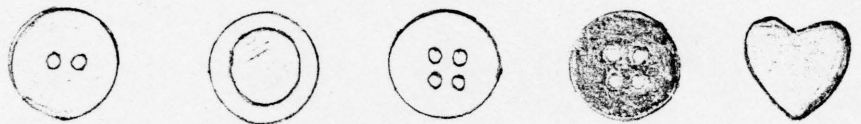
Options:



Item 7. Task photographed: Three buttons different size/color/
shape/number of holes/no holes.

Directions: The buttons go together for a reason. Dis-
cover the reason, find the button that
goes with this set for the same reason.

Options:



Item 8. Task photographed: Four buttons alternating size/shape/
color.

Directions: Discover the pattern, find the button that
goes next in line. Put it on the picture
next in line.

Options:



Item 9. Task photographed: Four buttons seriated S/L alternating color/number of holes.

Directions: Discover the pattern, find the button that goes next in line. Put it on the picture next in line.

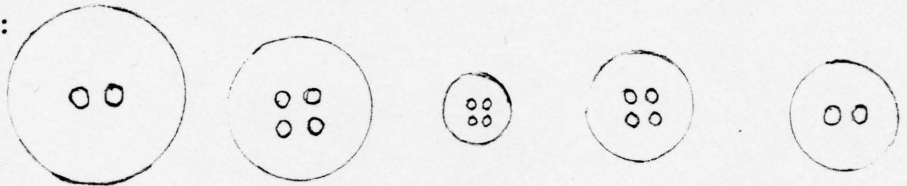
Options:



Item 10. Task photographed: Four buttons seriated S/L all same color alternating number of holes.

Directions: Discover the pattern, find the button that goes next in line. Put it on the picture next in line.

Options:



APPENDIX B

SAMPLE LETTER TO THE PARENTS

Dear Parents:

I am requesting consent for your child to participate in a dissertation study that is seeking to establish the effectiveness of an instrument to assess children's learning.

Children from the first grade rooms are being selected to participate in small group tests. The information gained will be shared with your child's teachers, but will otherwise be confidential. The children's names will be coded into number so that anonymity will be maintained in a general report that will be prepared. The procedures are those similar to the ones used in standardized test situations.

Should you desire further information on the project please feel free to call me at my office _____ to _____ or my home after _____. I am a parent of two _____ students and would be happy to respond to any questions.

Please complete the form below and return it today in the stamped, addressed envelope.

Thank you for your cooperation in this effort to find additional means of understanding children's learning process.

Sincerely,

_____ my (son) (daughter) has my permission to participate in the project directed by _____ to be conducted at _____ School.

Date

Signature of Parent or Guardian

LIST OF REFERENCES

- Aldrich, R. A. Marcy open school: Feeding back to decision makers. In V. Perrone, E. M. Cohen (eds.), Testing and evaluations: New views. Washington, D. C.: Association for Childhood Educational International, 1975.
- Anastasi, A. Psychological testing. (4th ed.) New York: Macmillan Publishing Co., Inc., 1976.
- Baratz, S. S., and J. C. Baratz. Early childhood intervention: The social science base of institutional racism. Harvard Educational Review, 1970, 40, 29-50.
- Beggs, D. L., and E. Lewis. Measurement and evaluation in the schools. Boston: Houghton Mifflin Co., 1975.
- Bersoff, D. Silk purses into sow's ears: The decline of psychological testing and a suggestion for its redemption. American Psychologist, 1973, 28, 892-899.
- Bortner, M. Review of progressive matrices. In O. K. Buros (ed.), The sixth mental measurements yearbook. Highland Park, N. J.: The Gryphon Press, 1965.
- Bruner, J. The process of education. New York: Vintage Books, 1963.
- _____ and J. Anglin (eds.). Beyond the information given. New York: W. W. Norton and Co., Inc., 1973.
- Buros, O. K. (ed.). The sixth mental measurement yearbook. Highland Park, N. J.: The Gryphon Press, 1965.
- Bussi, A. M., and E. A. Chittenden. Reflections in teaching. In V. Perrone and E. M. Cohen (eds.), Testing and evaluation: New views. Washington, D. C.: Association for Childhood Education International, 1975.
- Carini, P. F. The prospect school: Taking account of process. In V. Perrone and E. M. Cohen (eds.), Testing and evaluation: New views. Washington, D. C.: Association for Childhood Education International, 1975.

- Carver, R. P. Two dimensions of tests. American Psychologist, 1974, 29, 512-518.
- Cazden, C. B. (ed.). Language in early childhood education. Washington, D. C.: National Association for the Education of Young Children, 1972.
- _____. Concentrated vs. contrived encounters: Suggestions for language assessment in early childhood education. Paper presented at Seminar on Language and Learnings in Early Childhood, Leeds University, Institute of Education, January 8-9, 1974.
- Chase, C. I. Measurement for education evaluation. Menlo Park, Cal.: Addison-Wesley Publishing Co., 1974.
- Cronbach, L. J. Essentials of psychological testing. (3rd ed.) New York: Harper and Row, 1949.
- Davis, F. B. (ed.). Standards for educational and psychological tests. Washington, D. C.: American Psychological Association, 1974.
- Durost, W. N., H. H. Bixler, J. W. Wrightstone, G. A. Prescott, and I. H. Balow. Metropolitan achievement tests, primer, teachers handbook. New York: Harcourt-Brace-Jovanovich, 1971.
- Dyer, H. S. Testing little children: Some old problems in new settings. Paper presented at National Leadership Institute, Washington, D. C., October 7, 1971.
- Ebel, R. L. (ed.). Encyclopedia of educational research. (4th ed.) New York: Macmillan, 1969.
- Estes, W. Learning theory and intelligence. American Psychologist, 1974, 29, 740-749.
- Fehr, F. S. Critique of hereditarian accounts. In Science, heritability and IQ. Cambridge: Harvard Educational Review Reprint Series, 1969, 4, 40-49.
- Fischer, C. Intelligence contra IQ. A human science critique and alternative to the natural science approach to man. In K. Riegel (ed.), Intelligence: Alternative views of a paradigm. New York: S. Karger, 1973.
- Gordon, I. J. Studying the child in the school. New York: Macmillan 1966.

- Hein, G. E. Standardized testing: Reform is not enough. In V. Perrone and E. M. Cohen (eds.), Testing and evaluation: New views. Washington, D. C.: Association for Childhood Education International, 1975.
- Henderson, R., and J. Bergan. The cultural context of childhood. San Francisco: Chandler Publishing Co., 1976.
- Hill, W. F. Learning: A survey of psychological interpretations. San Francisco: Chandler Publishing Co., 1963.
- Horowitz, F. D., and L. Y. Paden. The effectiveness of environmental intervention programs. In B. M. Caldwell and N. Ricciuti (eds.), Review of child development research, child development and social policy. Chicago: The University of Chicago, 1973.
- Hughes, M. A tentative hierarchy of mental activity for heuristic purposes only. Unpublished manuscript, Arizona Center for Educational Research and Development, University of Arizona, 1967.
- Hunt, J. McV. Intelligence and experience. New York: The Ronald Press Co., 1961.
- Inhelder, B., and J. Piaget. The early growth of logic in the child. New York: Norton and Co., Inc., 1964.
- Jensen, A. R. How much can we boost IQ and scholastic achievement. In Environment, hereditary and intelligence. Cambridge: Harvard Educational Review Reprint Series, 1969, 2, 1-123.
- Labov, W. Academic ignorance and black intelligence. Atlantic Monthly, 1972, 229, 56-67.
- Levine, M. The academic achievement test. American Psychologist, 1976, 31, 228-237.
- Linn, R. L., and J. A. Slude. The determination of the significance of change between pre- and posttesting periods. Review of educational research, Winter 1977, 47, 121-150.
- Matarazzo, J. D. Wechsler's measurement and appraisal of adult intelligence. (5th ed.) Baltimore: The Williams and Wilkins Co., 1972.
- McCandless, B. R., and E. D. Evans. Children and youth: Psychosocial development. Hinsdale, Ill.: The Dryden Press, 1973.

- McCarthy, D. Manual for the McCarthy scales of Children's Abilities. New York: The Psychological Corp., 1972.
- McClelland, D. C. Testing for competence rather than for "intelligence." American Psychologist, 1973, 28, 1-14.
- Mehrens, W. A., and I. J. Lehmann. Standardized tests in education. New York: Holt, Rinehart and Winston, Inc., 1969, 40-41.
- Meier, D. Another look at what's wrong with reading tests. In V. Perrone and E. M. Cohen (eds.), Testing and evaluation: New views. Washington, D. C.: Association for Childhood Education International, 1975.
- Mercer, J. R. Sociocultural factors in the educational evaluation of black and chicano children. Paper presented at the 10th Annual Conference on Civil and Human Rights of Educators and Students, Washington, D. C., February 18-20, 1972.
- Meyers, E. S., H. H. Ball, and M. Crutchfield. The kindergarten teacher's handbook. Los Angeles: Gramercy Press, 1973.
- Murphy, L. B. The stranglehold of norms on the individual child. In V. Perrone and E. M. Cohen (eds.), Testing and evaluation: New views. Washington, D. C.: Association for Childhood Education International, 1975.
- Nunnally, J. C. Psychometric theory. McGraw-Hill Series in Psychology. New York: McGraw-Hill, Inc., 1967.
- Patton, M. Q. Understanding the gobble-dy-gook: A people's guide to standardized test results and statistics. In V. Perrone and E. M. Cohen (eds.), Testing and evaluation: New views. Washington, D. C.: Association for Childhood Education International, 1975.
- Paul, A., A. Smith, and R. Henderson. Intellectual kits: Tools for instruction in the Tucson Early Education Model. Unpublished manuscript, Arizona Center for Educational Research and Development, University of Arizona, 1970.
- Richards, J. M., Jr. A simulation study of the use of change measures to compare educational programs. American Educational Research Journal, 1975, 12, 299-311.
- Robb, G. P., L. C. Bernardoni, and R. W. Johnson. Assessment of individual mental ability. San Francisco: Intext Educational Publishers, 1972.

- Samuda, R. J. Psychological testing of American minorities: Issues and consequences. New York: Dodd, Mead and Co., 1975.
- Sarason, S. B. The unfortunate fate of Alfred Binet and school psychology. Teachers College Record, 1976, 77, 581-592.
- Saunders, T. F. Double Think. Tucson: Farmington Press, 1973.
- Sigel, I. E. The search for validity or the evaluator's nightmare. In R. A. Weinberg and G. Moore (eds.), Evaluation of educational programs for young children. Washington, D. C.: The Child Development Associate Consortium, 1975.
- Silberman, C. Crisis in the classroom. New York: Random House, 1970.
- Stinchcombe, A. L. Environment: The cumulation of events. In Science, heritability and IQ. Cambridge: Harvard Educational Review Reprint Series, 1969, 4, 28-39.
- Stodolsky, S. S. What tests do and don't do. In V. Perrone and E. M. Cohen (eds.), Testing and evaluation: New views. Washington, D. C.: Association for Childhood Education International, 1975.
- Thornburg, H. D. School learning and instruction. Monterey, Cal.: Books and Cole Publishing Co., 1973.