DOCUMENT RESUME

ED 365 520 SE 053 887

AUTHOR	Trombley, Robert J.; Weiss, David J.
TITLE	Measurement of Basic Skills in Mathematics.
INSTITUTION	National Center for Research in Vocational Education, Berkeley, CA.
SPONS AGENCY	Office of Vocational and Adult Education (ED), Washington, DC.
PUB DATE	Jan 93
CONTRACT	V051A80004-89A
NOTE	46p.
AVAILABLE FROM	National Center for Research in Vocational Education, Materials Distribution Service, Western Illinois University, 46 Horrabin Hall, Macomb, IL 61455 (Stock No. MDS-101).
PUB TYPE	Information Analyses (070)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Basic Skills; *Diagnostic Tests; Higher Education; Mathematics Achievement; Mathematics Education; *Mathematics Skills; Student Evaluation; *Student Placement; *Test Construction; *Vocational Education
IDENTIFIERS	Mathematics Education Research

ABSTRACT

This paper examines the process of developing a measuring instrument to measure basic skills in mathematics for those students entering post-secondary vocational education programs. The paper is organized around the steps in the process of developing such an instrument. The first section addresses the need to define conceptually the construct of basic mathematical skills. Three meanings of the concept of hierarchical development in mathematics are discussed: (1) a logical sequence, referring to the structure inherent within the topic; (2) a psychological sequence, referring to the order in which a topic can be learned; and (3) an instructional sequence, referring tc the order in which a topic is taught. The second section focuses on research on mathematical abilities and the development of models through the use of factor and hierarchical analysis. The third section, on test development, discusses the need for measures of basic skills in mathematics, the limitations on the usefulness of global sources in test development, and alternatives to global measurement for evaluating mathematical learning. The paper concludes that, in order to develop a diagnostic test battery for measuring mathematical ability, the domain structure must be clearly defined, continued research must clarify the nature of mathematical ability, and methods of providing maximal diagnostic capability in a minimal amount of student testing time must be developed. Contains 77 references. (MDH)

NCRVE

National Center for Research in Vocational Education

University of California, Berkeley

MEASUREMENT OF BASIC SKILLS IN MATHEMATICS

Supported by
the Office of Vocational and Adult Education,
U.S. Department of Education

This publication is available from the:

National Center for Research in Vocational Education
Materials Distribution Service
Western Illinois University
46 Horrabin Hall
Macomb, IL 61455

800-637-7652 (Toll Free)

MEASUREMENT OF BASIC SKILLS IN MATHEMATICS

Robert J. Trombley

David J. Weiss

University of Minnesota

**National Center for Research in Vocational Education
University of California at Berkeley
1995 University Avenue, Suite 375
Berkeley, CA 94704**

Supported by
The Office of Vocational and Adult Education,
U.S. Department of Education

January, 1993

MDS-101

FUNDING INFORMATION

Project Title:	National Center for Research in Vocational Education
Grant Number:	V051A80004-89A
Act under which Funds Administered:	Carl D. Perkins Vocational Education Act P.L. 98-524
Source of Grant:	Office of Vocational and Adult Education U.S. Department of Education Washington, DC 20202
Grantee:	The Regents of the University of California National Center for Research in Vocational Education 1995 University Avenue, Suite 375 Berkeley, CA 94704
Director:	Charles S. Benson
Percent of Total Grant Financed by Federal Money:	100%
Dollar Amount of Federal Funds for Grant:	$5,744,000
Disclaimer:	This publication was prepared pursuant to a grant with the Office of Vocational and Adult Education, U.S. Department of Education. Grantees undertaking such projects under government sponsorship are encouraged to express freely their judgement in professional and technical matters. Points of view or opinions do not, therefore, necessarily represent official U.S. Department of Education position or policy.
Discrimination:	Title VI of the Civil Rights Act of 1964 states: "No person in the United States shall, on the grounds of race, color, or national origin, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program or activity receiving federal financial assistance." Title IX of the Education Amendments of 1972 states: "No person in the United States shall, on the basis of sex, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any education program or activity receiving federal financial assistance." Therefore, the National Center for Research in Vocational Education project, like every program or activity receiving financial assistance from the U.S. Department of Education, must be operated in compliance with these laws.

MEASUREMENT OF BASIC SKILLS IN MATHEMATICS

Robert J. Trombley and David J. Weiss

University of Minnesota

The process of developing a measuring instrument to measure any educational or psychological variable requires that the domain be carefully delineated and specified (Linn, 1980). This process usually begins with a review of the conceptual literature on the topic. The next step is to evaluate the empirical research on the nature of the domain under consideration. A third step is to examine the literature on how others who have attempted to measure the domain have approached the measurement problem, evaluate any special problems that they encountered, and consider possible solutions to these problems. Once this process is completed, the development of the measuring instrument proceeds from the definition of the domain and the experiences of others who have attempted to develop measuring instruments for that domain. The development of the measuring instrument includes writing test items to sample the relevant aspects of the domain, and the selection and application of the relevant psychometric technology to implement the measurement process. The present paper follows this approach, with a focus on the measurement of basic skills in mathematics.

Assessment serves many purposes and it is becoming increasingly important in many aspects of everyday life (Haertel & Calfee, 1983). In education, for instance, tests and assessment batteries have been used for such diverse purposes as certifying competency for high school graduation, college placement, and in evaluating the effectiveness of school district policies (Baker & Herman, 1983). Hieronymous (1972) identified ten purposes for testing. Shoemaker (1975) observed that the list clustered into two groups--individual assessment and group assessment. While Hieronymous' list specifically concerns reading skills, the list--particularly those items which concern individual assessment--can be generalized to include the assessment of mathematical skills. Shepard (1980) suggests that there are three reasons for using tests. One is for classification of people, the second is for diagnosis of individuals, and the third is concerned with program evaluation.

Diagnostic tests, which are by definition individual assessment instruments, are used to determine whether someone is ready to move on to more advanced work, needs additional work at their current level, or needs remedial help. In educational institutions, for instance, diagnostic testing is used to evaluate students' progress with respect to a series of educational objectives by matching educational objectives with the test items (Shepard, 1980). Wood (1980) eloquently depicted the need to do diagnostic testing for deficiencies in mathematical ability in today's educational environment:

> Despite the current emphasis on mathematics in both elementary and secondary schools, modern high schools still offer varying routes to a diploma, not all of which prepare a student for college entrance. But a diploma from an accredited high school, however acquired, remains a sufficient condition for entrance in many colleges, particularly junior colleges. Colleges with an open-door policy are therefore faced with ever-larger numbers of entering freshmen who have studied very little or no mathematics (p. 59).

Generally, the domains used for diagnostic testing are narrowly defined because the interest of the examiner lies in whether a particular skill has been learned, rather than a global score which reflects the general attainment across a heterogeneous domain of skills. A clear, concise definition of the domain being tested is particularly important in the assessment of basic skills in mathematics; however, there has been some controversy among mathematicians as to the nature of "mathematics" (Rees, 1962; Freemont, 1969), and this has led to different categorizations and partitioning of mathematics (Griffiths & Howson, 1974). Changes in the way mathematics is viewed can affect the mathematics curriculum (Robitaille & Dirks, 1982). One such curriculum change that has taken place in the last twenty years is the introduction of "New Math" concepts. Instead of splitting mathematics into a great number of small domains, concepts such as set theory provide a unifying structure to the mathematics curriculum. Such debates and innovations underscore the need to develop a clear and concise description of the content domain prior to the development of measuring instruments.

In the remainder of this review a closer look is taken at the structure of the domain of mathematics as it pertains to the assessment of basic skills in mathematics. "Basic skills" means those skills that the majority of high school graduates would be able to perform successfully after exposure to the typical mathematics curriculum. Special emphasis is placed on the relationship of the content domain to the purposes for which assessment is being implemented. In particular, the efficacy of using global scores versus sub-scale scores for assessing the basic mathematics skills of students entering post-secondary vocational education programs is evaluated.

CONCEPTUAL DEFINITIONS OF MATHEMATICS SKILLS

Very (1967) suggests that math ability is a multifaceted construct that, in general, reflects the ability to do quantitative thinking, or, more specifically, to be able to "discover, manipulate, and evaluate relationships" (p. 172). Very noted that mathematicians whom he queried about the nature of mathematical ability did not provide any consistent definition of the subject, but for Very their responses suggest two general approaches to understanding mathematical ability. One group tended to match ability with individual courses like algebra, trigonometry, and calculus, while the other group linked ability to theoretical processes such as general reasoning.

Whether Very's conception of mathematical ability is accepted, or definitions of mathematical ability provided by mathematicians are used, there are three common elements to these conceptualizations of mathematical ability. Each view involves a content domain, an individual, and a cognitive process. In a similar vein, Hart (1981) pointed out that there are three meanings to the concept of hierarchical development in mathematics: (1) a logical sequence, which refers to the structure inherent within the topic; (2) a psychological sequence, which refers to the order in which a topic can be learned; and (3) an instructional sequence, which refers to the order in which a topic is taught.

Hart pointed out that while these three ways of understanding mathematical hierarchies are independent, they are all necessary to successfully learn how to do mathematics. This suggests that when a set of operational definitions in mathematics is being explicated for use in test development, the structure of mathematics must be considered from all three perspectives--logical sequence, developmental sequence, and curriculum sequence. This section reviews the literature concerning each of these structural definitions of mathematics.

The Logical Structure of Mathematics

It is difficult to singularly characterize the domain of mathematics. Spitznagel (1971) views mathematics in terms of the objects of mathematical inquiry. Mathematics, according to Spitznagel, is concerned with "quantitative, spatial, and logical relationships" (p. 2). Kline (1962) has said, "One can look at mathematics as a language, as a particular kind of logical structure, as a body of knowledge about number and space, as a series of methods for deriving conclusions, as the essence of our knowledge of the physical world, or merely as an amusing intellectual activity" (p. 2). Another characteristic of mathematics is that it does not "deal with particular things or particular properties: we [mathematicians] deal formally with what can be said about *any* thing or *any* property" (Russell, 1919, p. 196, italics in original). Kline (1962), evaluating the developmental history of mathematics, graphically illustrates Russell's observation. Kline suggests that people first thought of numbers in terms of objects. Later they began to abstract the numbers from the objects themselves so that they could think about whole numbers without attaching them to specific objects like apples or sheep. The process of discovering mathematical concepts is both developmental and hierarchical. As Kline stated, "On the basis of elementary abstractions, mathematics creates others which are even more remote from anything real. Negative numbers, equations involving unknowns, formulas and other concepts ... are abstractions built upon abstractions" (p. 31).

Mathematics is a very large field and its structure is difficult to describe. There are about one hundred recognized subdisciplines in mathematics; if specialized fields of mathematics, such as theoretical physics and operations research, are included the number of disciplines would be several hundred (Steen, 1978). In point of fact, the domain of mathematics is so large that it must be divided into subdomains in order to grasp its nature and extent, even if the subdomain boundaries are not exactly defined (Stein, 1963). One way of coping with such a large domain is to break the field down into a few very broad categories. According to Kothe and Ballier, the work of Bourbaki, a group of French mathematicians who have been publishing a multi-volume work titled *Ele'ments de Mathe'matique*, which treats mathematics as the theory of structures, consists of reorganizing mathematics by selecting worthwhile structures and "arranging the various structures according to their mutual relationships [by] incorporating them in a natural way into the edifice as a whole, which then may be called a hierarchy of structures" (p. 521). The three primary structures that Bourbaki adopted are algebraic structures, topological structures, and ordered sets. Similarly, Steen (1978) refers to algebra, analysis, and topology as representative of the "common culture of modern mathematician[s]" (p. 6). Other authors, however, do not emphasize fundamental branches of mathematics; rather, they treat each major area of mathematics as separate branches (see Kline, 1962; and Stein, 1963).

Although fundamental branches in mathematics can be considered independent, the structure within a branch is hypothesized to be hierarchical. As Kothe and Ballier (1974) stated, "the whole edifice of algebra, of general topology and of the theory of ordered sets will therefore be constructed by first investigating the most general of these structures, and then proceeding to more special structures by the stepwise adjunction of further axioms" (p. 522). A similar concept is suggested by Kline (1962). Kline argued that each branch, although distinct, has the same fundamental logical structure. For instance, each begins with a fundamental concept such as whole numbers in the mathematics of numbers, or the concepts of point, line, and triangle in Euclidean geometry. The concepts also obey explicitly stated axioms, and theorems derived from concepts and axioms. For example, Kline argued that "Arithmetic, algebra, the study of functions, the calculus, differential equations, and various other subjects which follow the calculus in logical order are all developments of the real number system" (p. 660).

Taken together, these conceptual definitions of the logical structure of mathematics suggest that the domain of mathematics can be represented as a collection of independent branches of specialized areas, and that within these separate areas there is a hierarchical structure. Although useful in a conceptual sense, these definitions of the domain of mathematics are not specific enough for the purposes of developing measurement instruments for measuring basic skills.

Developmental Structures

The second type of hierarchical structure, the developmental process, suggested by Hart (1981) concerns the order in which material can be learned. The learning process involves the integration of the domain, the purpose for which it is being learned, and the student (Dienes, 1960). According to Bloom, Hastings, & Madaus (1971) the learning process, which is driven by pedagogical considerations, can be facilitated by the intrinsic structure of the domain. Three models which typify the cognitive developmental models are the formalistic model, the structuralist model, and the learning model.

The central concept of Piaget's formalistic model of development is that children go through a series of cognitive stages where each successive stage represents a higher level of cognitive development (see Piaget, 1952). The first stage is represented by a dominance of sensory perceptions and an inability to mentally reconstruct a prior situation. At the second stage, which Piaget labeled preoperational, the child is dependent on his/her immediate perceptions. The third stage is one of concrete operations. Finally, at about the age of adolescence children enter the stage of formal operations. This stage is characterized by the beginnings of advanced mathematical and scientific reasoning (Resnick & Ford, 1981).

There are several systems based on a structuralist model (Griffiths & Howson, 1974; Resnick & Ford, 1981). In general, these developmental theories maintain that the complex structures in mathematics can be broken down into simple structures which can be taught to children of any age. Development occurs as children first learn these simple structures and then later combine the simple structures into more and more complex structures. Thus, development is a process that incorporates both learning the structure and the process by which structures are related (Griffiths & Howson, 1974). Bruner (1960) suggests that mathematical development occurs in children as they "discover" the structure of mathematics. The child's intuition allows him/her to grasp the meaning of the structure even though they are not yet able to articulate, in a formal way, what it is they have discovered. The process begins to resemble a spiral curriculum where ideas are first presented at one level and then returned to at a later stage to be developed further.

A related developmental structure is a learning model devised by Dienes (1973). In Dienes' system a child progress through six stages as he/she progresses from being totally unaware of mathematics to understanding the theorems that underlie mathematics. In the first stage, the child comes to understand and adapt to a learning environment. The environment contains the tools and examples of the mathematical concept which the child will eventually learn. In the second stage, the child learns the rules which govern the situation. During the third stage, the child learns to abstract common elements from different concepts that have the same underlying structure. The fourth stage is characterized by the development of a symbolic system to represent the abstraction. This allows the child to talk about the abstraction itself without reference to the particular concepts from which the abstraction was derived. During the fifth stage the representation is examined in order to understand its properties. This process requires that a language be created as part of the descriptive process. These descriptions can later become axioms or even theorems. Finally, in the sixth stage the descriptions are bounded into a finite domain and the rules for moving through the domain are specified.

The three models of the development of mathematical ability outlined above suggest that achievement in mathematics depends on the person's state of cognitive development as well as his/her familiarity with the content domain being assessed. Two criticisms of this model can be raised, however. One is that all three models are dependent on age. For instance, Piaget's model suggests that most people will develop stage four mathematical processing abilities in their early teens. If this is correct, then evaluating this dimension of mathematical ability will not provide much information to aid in assessing a person's ability if they are over a certain age or if the goal of assessment is not limited to a specific stage of development, such as basic skills like computational ability.

A second area of concern is whether these models represent a global evaluation of the person or whether the evaluations are specific to each branch of mathematics being measured. In the former case the value of the evaluation is limited because of the age effects; however, a cognitive development profile score that was area specific could be quite useful with any age group if the level of cognitive development for individuals varies across branches of mathematics (see Kolen & Jarjoura, 1984). The models reviewed here, standing alone, do not however provide sufficient operational definitions for designing a measuring instrument which would measure cognitive development across branches of mathematics.

Curriculum

Bloom, Hastings, and Madaus (1971) stated that the primary function of education is the development of individuals so that they are "able to live effectively in a complex society" (p. 6). This naturally leads to a system where the goals of teaching, which are based on the sequence of curriculum being taught, are objectified so that students' progress toward these goals can be quantified. In Bloom's taxometric design, student learning is broken down into a set of behavioral objectives. In rough terms, an objective represents the way a teacher expects the behavior of a student to be changed. Furthermore, the behavioral objective should be clearly defined so that any other teacher would be able to determine whether the objective has been reached. The individual cells of a rectangular matrix, where behavioral objectives are listed along the horizontal axis and subject content along the vertical axis, represents specific content in relation to a particular objective or behavior.

Wilson (1971) developed a model for evaluating mathematics achievement based on the guidelines proposed by Bloom. This model concerned only grades 7 through 12, however, since the mathematics curriculum is sequential from kindergarten on up, it can easily be extended to the elementary grades. On the vertical, content dimension Wilson includes number systems, algebra, and geometry as general classifications. Each category is broken down into several subcategories (see Table 1). The behavior dimension, in this case cognitive behaviors, is subdivided into four categories--computation, comprehension, applications, and analysis. Just as with the content dimension, each cognitive behavior is capable of being further subdivided into more specific behaviors (see Table 1). The cognitive dimension is arranged based on a hierarchy of cognitive complexity and not just the difficulty of the task. It is hierarchical in that an item representing an "application" would also require the student to select the appropriate "operation" (comprehension) and be able to perform the "calculation" (computation).

The terms used in Table 1 for the content dimension are rather straightforward, but the terms used for the cognitive dimension require further elaboration. Computation items emphasize knowing how to perform particular operations and, of course, being able to perform these operations. Doing square roots or conversion of fractions to decimals are good examples of this computational behavior. Comprehension items

Table 1
The Content and Cognitive Dimensions of Wilson's (1971)
Table of Specifications for Secondary School Mathematics

Content Dimension

- 1.0 Number systems
  - 1.1 Whole numbers
  - 1.2 Integers
  - 1.3 Rational numbers
  - 1.4 Real numbers
  - 1.5 Complex numbers
  - 1.6 Finite number systems
  - 1.7 Matrices and determinants
  - 1.8 Probability
  - 1.9 Numeration systems
- 2.0 Algebra
  - 2.1 Algebraic
  - 2.2 Algebraic sentences and their solutions
  - 2.3 Relations and functions
- 3.0 Geometry
  - 3.1 Measurement
  - 3.2 Geometric phenomena
  - 3.3 Formal reasoning
  - 3.4 Coordinate systems and graphs

Cognitive Dimension

- A.0 Computation
  - A.1 Knowledge of specific facts
  - A.2 Knowledge of terminology
  - A.3 Ability to carry out algorithms
- B.0 Comprehension
  - B.1 Knowledge of concepts
  - B.2 Knowledge of principles, rules, and generalizations
  - B.3 Knowledge of mathematical structure
  - B.4 Ability to transfer problem elements from one mode to another
  - B.5 Ability to follow a line of reasoning
  - B.6 Ability to read and interpret a problem
- C.0 Application
  - C.1 Ability to solve routine problems
  - C.2 Ability to make comparisons
  - C.3 Ability to analyze data
  - C.4 Ability to recognize patterns, isomorphisms, and symmetrics
- D.0 Analysis
  - D.1 Ability to solve nonroutine problems
  - D.2 Ability to discover relationships
  - D.3 Ability to construct proofs
  - D.4 Ability to criticize proofs
  - D.5 Ability to formulate and validate generalizations

require that the student understand mathematical concepts and be able to generalize them. Application items require the student to be able to determine for any particular situation what specific mathematics concepts are required to arrive at a solution and, of course, to be able to perform these operations. Finally, analytical items concern the highest order of cognitive processing. This may mean the non-routine application of concepts or developing proofs.

The basic two-dimensional model developed by Wilson (1971) has been used in a number of curriculum studies. In a comparative study of national curriculua (Garden & Robitaille, 1989), the International Association for the Evaluation of Educational Achievement used a testing grid very similar to that developed by Wilson. The cognitive behavior dimension depicted by Wilson was adopted in its entirety; the content dimension was first divided into five strands--Arithmetic, Algebra, Geometry, Descriptive Statistics, and Measurement--and then subdivided into 133 categories. A very similar model was used in the National Longitudinal Study of Mathematical Abilities (Howson, Keitel, & Kilpatrick, 1981) study. According to Howson, et al., "The designers of NLSMA believed that achievement in mathematics is multifaceted and that its assessment required a battery of short tests . . . aimed at different facets" (p. 189). Wilson's model is ideally suited for this purpose because the individual cells of its two-dimensional matrix provide a constrained description of the domain from which test items can be developed.

Bruner (1960) pointed out that one key element necessary for students to learn new concepts is the sequence in which the information is presented. Indeed, the integration of mathematics depends upon building upon prior years of instruction. One area of research which has directly considered sequencing is curriculum development. Curriculum sequencing refers to the ordering of a structured domain consistent with the methods used for instruction, although it is not necessarily true that the sequential structure developed for instruction will mirror the content domain. Bloom, Hastings, and Madaus (1971) pointed out that if instruction in mathematics is to be successful, the curriculum needs to be sequenced. Because mathematics has both independent branches and a hierarchical structure within branches, sequencing can be a difficult task.

Strict sequencing of curriculum tends to follow a hierarchy based on difficulty. As Resnick and Ford pointed out ". . . the school curriculum has classically followed a path from simple to complex--from adding and subtracting single digits to computing complex multidigit arithmetic problems to solving algebraic equations." (p. 38) The same attitude toward the importance of sequencing is expressed by those who are directly involved in setting curriculum. The following expression on the purpose and importance of sequencing is taken from "Mathematics Activities, Level Nine" (1982):

> The mathematics curriculum for grades 1-8 is divided into 20 teaching levels. This organization is designed to improve instruction and to foster the continuous development of children. Each level embodies carefully delineated areas of learning arranged in progressive stages. Such an arrangement of sequential skills and subject matter eliminates grade restrictions and permits continuous growth according to the individual's ability and rate of learning (p. V.)

The origins of curriculum models can be found in both the theoretical investigations of the structure of mathematics and in basic research into the process of cognitive development. A model such as that offered by Wilson, represents a synthesis of these two independent lines of investigation. In particular, Wilson's model provides a structure matrix of target behaviors which can be used to design instructional sequences. Furthermore, the behavioral objectives matrix developed by Wilson provides a structure sufficiently detailed to serve as a starting point for the development of a basic skills test battery.

RESEARCH ON MATHEMATICAL ABILITIES

Guion (1977) suggested that in order to establish content validity the domain must represent behavior with a "generally accepted meaning" (p. 6). Empirical investigations of the observed behaviors that reflect elements of the domain provide one type of evidence validating the operational definitions used to develop a measuring instrument. These investigations must be undertaken whether the practitioner wishes to evaluate a person's score on either the content dimension or the cognitive dimension, or a mixture of the two such as Wilson's two-dimensional matrix of behavior objectives.

There are two primary lines of investigation in mathematical ability assessment--factor analyses of mathematical ability and studies of mathematical hierarchies. Factor analytic studies (1) have studied mathematical ability in relationship to other abilities such as verbal ability, and (2) have looked at mathematical ability in isolation by factoring only mathematical items. Studies of mathematical hierarchies have generally focused on circumscribed aspects of mathematics, such as the hierarchical structure of addition skills. This section reviews literature focusing on these two areas of empirical research.

Factor Analysis of Mathematical Abilities

Factor analysis, according to Harman (1976), "provides mathematical models for the explanation of psychological theories of human ability and behavior" (p. 3). The intent of factor analysis is to reduce the original set of variables into a smaller number of "factors" that retain the common information contained in the original variables, but which are more interpretable. There are a number of factor analytic models available, and any particular model may represent particular psychological theories better than another. In addition, decisions about what type of factoring method to use or what criterion to use for rotation can affect the conclusions derived from an empirical factor study. As a consequence of these multiple decisions, which must be made during the course of a factor analytic study, the interpretation of factor analytic studies involves some level of subjective decision making.

One of the key debates that has occurred in factor analytic work is concern for the nature of "ability." Spearman (1904) proposed a model for ability which emphasized a general factor. Spearman's "general factor" (*g*) represents the general cognitive ability of people across a wide range of separate abilities. In Spearman's approach, specific abilities, such as verbal and numerical, appear as narrowly defined factors which are specific to the type of operation being tested.

An alternative model postulates a series of specific abilities which are to a high degree independent. Thurstone's investigations into primary mental abilities are a good example of studies which consider ability to be made up not of "*g*," but rather a constellation of specific abilities. Thurstone's (1938) investigation of primary abilities identified a cognitive factor on which eight variables having high to moderate loadings measured numerical ability. Thurstone called this factor "N" for numerical. This factor included the four basic arithmetic operations--addition, subtraction, multiplication, and division--which had loadings (correlations with the factor) between .62 (division) and .81 (multiplication). Other tests, such as arithmetic reasoning and numerical judgment had loadings of .38 and .43 respectively.

After Thurstone's investigations into specific mental factors during the 1930s, a series of factor analytic studies of mathematical ability were undertaken in the 1940s and 1950s. These studies further investigated the nature of mathematical ability. In a study of numerical ability, Coombs (1941) factor analyzed a battery of tests that included separate tests of two-, three-, and four-digit addition and multiplication. Perceptual speed, verbal ability and other tests designed to assess various hypotheses about the nature of numerical ability were also included. The first factor extracted was labeled a "number" factor by Coombs. The two-, three- and four-digit addition problems loaded .74, .72 and .66 on it and the multiplication test loaded .64. These variables did not load highly on any of the other factors such as verbal, space, or perceptual speed. Coombs also extracted a deductive factor on which arithmetic and number series had the highest loadings (.57 and .56 respectively) and an inductive factor on which addition loaded .32.

Fruchter (1952) factored the subtests of the Airman Classification Test Battery along with the Guilford-Zimmerman Aptitude Survey, the AF Aircrew Classification Test Battery, the Army Classification Test Battery, and a battery of temperament and intelligence tests. The first factor extracted after rotation to simple structure was a verbal factor and the second factor "represents the ability to do arithmetical computations speedily and accurately" (p. 31). Subtests that had high loadings on the second factor included the Gray-Votaw Arithmetic test and the Differential Aptitude Test of numerical ability, both of which contain arithmetic computation items. The Aircrew Classification Test Battery Numerical Operations II test, which includes subtraction and division items, and the Army Classification Test Battery Arithmetic Computation test also had high loadings on this factor.

A series of factorial studies of mathematical ability done in the early 1950s by Guilford and his associates provided additional evidence for a numerical factor and several more specialized mathematical ability factors (Northrop, 1977). Results reported by Green, Russell, Guilford, and Christensen (1953) typify these studies. Green et al. factor analyzed 34 tests including many that were used by Fruchter (1952). After rotation, using the Zimmerman graphic orthogonal system of rotation, a verbal comprehension factor accounted for the most variance and a numerical facility factor was the second most important factor. The test which had the highest loading on the numeric facility factor was numeric operations, which included addition, subtraction, multiplication, and division problems.

The purpose of Green's et al.'s research was to identify factors that could be labeled as "reasoning." Two factors were extracted on which tests that assessed various aspects of reasoning ability had high loadings. In particular, for the first reasoning factor (which was labeled as a general reasoning factor), problem solving (which measures the ability to solve arithmetic-reasoning problems), and symbol manipulation had the highest loadings (.43 by one researcher and .53 by a second). Problem solving also had a moderate loading on the second reasoning factor, logical reasoning. Several subtests of mathematical ability such as number and operation changes (equation operations) loaded on several reasoning factors, which suggests that certain mathematical procedures are complex in nature and provide information about a number of distinct abilities.

In a study of gender differences in the factor structure of mathematics ability, Very (1967) found number facility to be the second factor extracted for both males and females. Consistent with the findings of Green et al. (1953), the tests that had the highest loading on this factor were addition, subtraction, multiplication, and division (.79 to .89). For women--but not men--one of two mathematical achievement tests used had a moderate (.31) loading on this factor, as did two tests of mathematical reasoning. For men, factor 5 represented an arithmetic reasoning ability. The two mathematics achievement tests loaded most highly on this factor, as well as all tests considered to represent arithmetic reasoning. The author suggested that incomplete factoring may explain why an arithmetic reasoning factor was not extracted for women.

Wrigley (1958) investigated the question of whether "mathematics should be regarded as an integrated whole" (p. 66) as measured by a special mathematical group factor, or whether the linkages of branches of mathematics (arithmetic, algebra, and geometry) can be adequately accounted for by a general factor which includes verbal ability. Based on his review of the literature from the turn of the century to the late 1950s, he determined that there was a split in the conclusions of prior researchers. According to Wrigley, research from both sides of the issue, however, suffered serious methodological flaws .

Wrigley's (1958) experimental design used measured mathematical ability based on problems measuring mathematical "attainment," as opposed to using aptitude items or a combination of both attainment and aptitude. Wrigley noted that he ran the risk of deriving a pedagogical factor because "what is measured is greatly dependent upon the teaching which children have received" (p. 69), but practical considerations about what can be done in a single experiment led him to select items based on a survey of course syllabi (p. 70). In addition to mathematics tests, several tests measuring a "g" factor were included. Factor analysis was done using centroid analysis.

As expected, the first factor to emerge, which accounted for 31% of the common variance, was a general factor. However, since the common variance accounted for only 61.5% of the total variance, the first factor accounted for only 19.3% of the total variance. The mathematical tests and the Manchester General Ability test had high loadings on the first factor, suggesting that mathematical ability is closely connected with general intelligence. The mathematical group factor, which accounted for 3.8% of the total variance, was the fourth factor extracted, and--as expected--the mathematical tests had their highest loadings on this factor.

The above factor studies of mathematical ability demonstrate that in a test battery that includes items with mathematical content, a general mathematical ability factor can be extracted. The nature of the mathematical ability factor is, however, difficult to characterize. Evidence suggests that the first mathematical factor does represent a general mathematical ability factor. Across the studies reported, numerous diverse mathematical tests have loaded on this factor, suggesting that the factor is tapping an ability which apparently is common to many areas of mathematics. On the other hand, these studies also show several lower-order factors which apparently relate to specific areas of mathematical ability. It is difficult to interpret these lower-order factors because the tests used in these studies often represent both a content dimension and a cognitive development dimension. In addition, the overall design of these studies perhaps led to a general inability to interpret lower-order mathematical factors, because the mathematics tests represented only a portion of the total number of tests factor analyzed; where the number of tests representing the construct of interest is small, little confidence can be placed in interpreting lower-order factors.

Wrigley's (1958) research was based on the factor analysis of tests; Furneaux and Rees (1978), however, suggest that research on the structure of mathematical ability should be based on the correlations among test items. Their 1978 study was a partial replication of a prior study, but a larger item set was used. The item set consisted of items that measured various mathematical operations. Factoring was done by maximum likelihood and Varimax rotation was used. Twenty-three factors, which accounted for only 49.6% of the total variance, were extracted. Six factors with eigenvalues over 1.0 were retained. The items which loaded on the first two factors had no obvious content or process in common; consequently, they were interpreted as general factors. They accounted for 22.5% and 16.5% of the common variance or 11.1% and 8.2% of the total variance, respectively.

Furneaux and Rees (1978) equated the second factor with a general intellectual ability factor, since the four Thurstone Primary Mental Abilities variables also loaded on this factor. The first factor to emerge contained a large number of "core" items. By this the authors seemed to mean items that were designed to be more difficult, or at least turned out to be more difficult, than expected. The authors call this factor an "inference" factor and they suggest that it represents a general mathematics ability factor.

Based on their research results, Furneaux and Rees suggest that "*g*" items are clearly structured problems which can be solved once the proper algorithms have been learned. Inference items, on the other hand, require the "ability to conceptualize the problem in such a way that the relevant operations can first be identified, and then applied in proper combination and sequence" (p. 512). It should be emphasized, however, that the first two factors accounted for only 40% of the common factor variance. The next four factors accounted for another 16.1% of the common variance, which means that the factors beyond the sixth accounted for approximately 44% of the common factor variance and 21.8% of the total variance. Indeed, Furneaux and Rees pointed out that 28 of the 69 mathematics items loaded only on factors beyond the sixth, thus in this study these items did not provide any significant information about a mathematical ability trait even though items that loaded on these separate, minor factors had either similar content or relied on similar processes.

Powers, Swinton, Spencer, and Carlson (1977) investigated the structure of the Graduate Record Exam. Twelve principal factors were extracted, which accounted for virtually all of the common variance but only about 40% of the total variance. After inspection of the roots, the first eight factors were retained. These factors accounted for 94% of the common variance and 37.5% of the total variance. The first three factors were taken to represent three global skills, two of which represented the two verbal sections and one representing the quantitative section. Little content classification covariance was found among verbal items, but indications of content structure among the quantitative items were found.

The first extracted factor was the quantitative factor, and it accounted for 28.3% of the common variance or 10.6% of the total variance. Forty-three of the fifty-five items in the quantitative section loaded on this factor, and none of the verbal items did. Three of the factors that we were extracted after the third global factor represented algebra, data interpretation, and application or word problems. Algebraic items loaded most heavily on factor 4, accounting for 7.5% of the common variance. Factor 6, accounting for 4.3% of the variance, was characterized by the data interpretation items, and factor 7 was best described as representing word problems. The structure for the smaller factors was not the same when the second form was analyzed. The first general factor to emerge was again a general quantitative factor, but of the three mathematical subgroups defined above only data interpretation emerged as a separate factor, and even it was split between two factors. Algebra and application items loaded on the general factor.

Martin and Dunbar (1985) factor analyzed the Iowa Tests of Basic Skills (ITBS) to determine whether, in addition to a general ability factor, secondary group factors existed. To increase the likelihood of discovering hypothesized group factors, the number of variables factor analyzed was increased by creating composites from subtest items. As expected, there was a very large general factor, but there were also four interpretable group factors. The fourth factor to be extracted, mathematics computation, had the highest subtest loadings of any of the composite variables. It is interesting to note that two of the computational composites also had moderate loadings on the sixth factor, although this factor was not interpreted by the authors. The authors suggested that these results do not support Klein's (1981) conclusion that the subscale scores do not provide information over and above what is generated by the ITBS total score.

In the Lawrence and Dorans (1987) study, items were prepared for factor analysis by grouping them into parcels which represented content domains (arithmetic, algebra, geometry, and miscellaneous) and similar item difficulties. The authors, using LISREL's confirmatory factor analysis capabilities, concluded that the one-factor solution provided a good fit to the item parcel data. The two-factor model, which assumed geometry items as forming the second factor, and the three-factor model, which allowed each content area to be represented by a factor, added very little to the fit of the data to the model. According to the authors, the results demonstrate that the SAT-Mathematical is unidimensional. For the three-factor model, the factors correlated above .92 across four SAT administration. The general factor accounted for 99% of the algebra parcel variance, 98% of the arithmetic parcel variance, and 90% of the geometry parcel variance. The authors concluded that the empirical evidence does not support reporting content area scores for SAT-Mathematical.

Reckase, Davey, and Ackerman (1989) investigated the structural properties of the Mathematics Usage Test (AAP Math), which is used to measure mathematics achievement of high school students in the content found in courses offered in grades seven through eleven. These tests were constructed using a set of content specifications; therefore, the tests may not be unidimensional either within a specific form or across several forms, and different forms may measure different dimensional structures. The content areas investigated included arithmetic and algebraic operations, arithmetic and algebraic reasoning, geometry, intermediate algebra, number and numeration concepts, and advanced topics. The analysis was performed on a six-by-six matrix of Pearson product-moment correlations of number-correct scores for each content area. All forms had a dominant first factor which accounted for 88% or more of the total variance, which suggest that AAP Math is highly unidimensional. Factor analysis done on the tetrachoric correlations suggested that a two-factor solution was appropriate. A multidimensional item response theory analysis was performed on each of the test forms. This analysis distinguished between computational items and word problem items, but these constructs were highly intercorrelated.

Summary

The above group of factor analytic studies of mathematical ability suggest that when mathematical ability is measured without reference to other abilities there is a relatively clear-cut structure to mathematical ability. In particular, when the studies use a heterogeneous sample, a general factor representing mathematical ability will emerge as a dominant factor. On the other hand, the studies generally point to specific area factors in mathematics which, while not as important as the general factor, account for a large amount of the total variation in the correlation matrix but only a small amount of the common variance. For designing a measurement instrument for mathematical skills, the usefulness of a general factor is limited, but information concerning specific mathematical factors is potentially useful when the descriptions of the specific factors can be used as operational definitions of the content domains being assessed.

Studies of Mathematical Hierarchies

It was noted above that the domain of mathematics can be conceived as composed of independent branches and hierarchical structures within branches of mathematics. The factor studies of mathematical ability support the conclusion that branches of mathematics can be treated as independent subdomains. Support for the nature of the relationships of mathematical concepts within a branch of mathematics requires studies which directly evaluate the structure of an area of mathematics.

One way to examine the relationship of concepts within a branch of mathematics is as a hierarchical structure. Learning psychologists view hierarchies as a transfer relationship between different tasks, where two tasks are in a hierarchical relationship if (1) one task is easier to learn than the other, and (2) if the simpler task is learned first. When such a relationship exists, it will be easier for a student to learn the more complex task having first mastered the easier task (Resnick, 1973)--learning how to add will make learning how to multiply easier, and being able to multiply will facilitate learning how to divide. Resnick (1973) refers to this as "a hierarchically organized *sequence* of tasks" (p. 312, italics in original).

Gagne and Paradise (1961) investigated the hierarchical structure of learning sets hypothesized as necessary steps in learning how to solve simple algebraic equations. Learning sets, according to Gagne and Paradise, mediate the positive transfer of knowledge between lower- and higher-order tasks. Knowledge of lower-order skills necessary to solve equations were determined by asking the question "What would an individual have to know how to do in order to achieve successful performance of this class of task, assuming he were given only instructions?" (p. 4). By applying this analysis to successively lower-order tasks, the learning sets become increasingly simple and increasingly general.

According to Gagne and Paradise there are four possible pass-fail relationships between any two learning sets. An examinee could pass both the lower level and higher level learning sets (+,+) or the examinee could fail at both tasks (-,-) or the examinee could pass the lower level set but fail the higher level set (+,-). These three relationships are consistent with what would be expected if there is a hierarchical structure to the learning sets. The mixed (+,-) relationship can occur because of an ineffective learning process. On the other hand, the other mixed pattern (-,+) is inconsistent with a hierarchical model. In the Gagne and Paradise (1961) study, 13.1% of the tested relationships showed this inconsistent pattern, 84.2% were consistent with the hierarchical model (-,- or +,+) and 2.7% neither supported nor refuted the model (+,-).

Gagne, Mayor, Garstens, and Paradise (1962) investigated the hierarchical structure of learning sets necessary to be able to do integer arithmetic. The results from this study generally support the hypothesized hierarchical structure. Eighty-two percent of the tested relationships were consistent with the hierarchical structure, but 17% of the relationships were found to be exceptions (-,+). In another study Gagne (1962) investigated the hierarchical structure of learning sets leading to the ability to derive formulas for the sum of n terms in a number series. For the seven students involved in this study, the hierarchical structure of the nine learning sets is clearly evident and, in this study, there were no exceptions (-,+) to the anticipated hierarchical relationship between learning sets. As can be seen in these three studies by Gagne and his associates, the existence of a hierarchically structure in specific areas of mathematics is supported. Two of the studies reported response patterns that are inconsistent with the proposed hierarchy. Gagne and Paradise (1961) suggest that unreliability in the measuring instrument may account for the inconsistent data.

White (1973) also considered reasons that would account for the inconsistent patterns, and listed four possible explanations: (1) there may be measurement error, since only one item was used to measure each element; (2) lower level skills may have been forgotten without affecting the higher level skills; (3) errors in the hierarchical structure; or (4) complete failure of the hierarchy model. The fact that 84.1% of the tested relationships in one study and 82% of the tested relationships in the other study supported the proposed hierarchical structure suggests that there is a reasonably well-defined hierarchy, but more research is needed to demonstrate which of the other three explanations--singly or in combination--would account for the response patterns that were inconsistent with the proposed hierarchy.

Airasian and Bart (1975) reanalyzed the data from Gagne et al.'s (1962) study of the hierarchy of addition skills using order theory (see Bart & Krus, 1973). Order theory circumvents the need to define each testable relationship as was done by Gagne et al (1962) and allows for the testing of logical equivalence and logical independence, in addition to prerequisite relationships. The hierarchical structure generated in this study was similar to that found in the Gagne et al. study, but was somewhat more complex. In particular, many indirect relationships were found that were not identified by Gagne et al.

Using Gagne and Paradise's (1961) method for constructing skill hierarchies, Linke (1975) constructed a learning hierarchy of graphical interpretation skills. This structure was far more complex than those investigated by Gagne. It was composed of 22 basic skills and six terminal skills. With certain minor exceptions, the predicted hierarchy was validated. Furthermore, the results were shown to be consistent with results from an independent replication. The inconsistent results, according to Linke, were likely due to incidental acquisition caused by redundancy in lower level skills acquired in learning different higher level skills.

Kolb (1967) hypothesized that a hierarchically based mathematical learning sequence would facilitate the acquisition of quantitative science abilities, whereas an unrelated learning sequence would not facilitate learning quantitative science skills. Kolb developed two exercises in static friction to represent scientific behaviors. The first contained exercises for generating operational definitions, and the second exercise provided experience in data interpretation. The science behaviors also included three quantitative performances which were then used as the final tasks of the mathematical learning hierarchy. The instructional sequence developed to support the final skills included topics in ratio and line segment graphs.

There were 26 tasks in Kolb's mathematical hierarchy. Using the ratio of positive transfer, with a critical value of .90 being necessary to conclude that a hierarchy of mathematical tasks existed, 25% of the hypothesized relationships failed to meet this criterion. However, it should be noted that all but one of these ratios were equal to or greater than .74, and three out of the seven that failed to make this criterion were equal to or greater than .85. Even though Kolb could not claim to have validated a specific mathematical hierarchy, there was strong support for his hypothesis that a curriculum designed according to a mathematical hierarchy facilitates the learning of science objectives.

Shermis (1988) investigated the relationship between undergraduate mathematics curriculum sequence and the difficulty of items that were selected to test learning of the skills in that curriculum. Principal components factor analysis was used to determine whether the data met the unidimensional assumption of one-dimensional item response theory (IRT) models, and IRT was used to determine item difficulties. Nearly 70% of the common variance was accounted for by the first factor, and the remainder was accounted for by the other two factors. It must be noted, however, that these three factors only accounted for 16.6% of the total test variance. As a consequence, the first factor accounted for only 11.6% of the total variance in the test items. Shermis noted that there was a high correspondence between the item rank--the position of the item in the curriculum sequence--and the item difficulty (Spearman's rho = .62, $p < .001$).

Summary

The results of the hierarchical studies reported here strongly suggest that, within narrowly defined areas of mathematics, there are hierarchical relationships between different mathematical processes. This research is limited in two ways, however. One limitation is that the mathematical hierarchies that have been researched and validated generally focus on a very narrowly defined area of mathematics. Further research needs to be done to show that the separately validated hierarchies may be linked together into a single hierarchical structure. A second limitation is that Gagne's procedure for constructing mathematical hierarchies to match a specific final task is too cumbersome to use when the hierarchy of relationships cover a very broad area (i.e., the hierarchical structure of arithmetic). Gagne's procedure may not be flexible enough to uncover a hierarchical relationships when the tasks are defined in this way.

One approach to task analysis suggested by Baker (1983), which may prove to be more flexible for working with broadly defined mathematical hierarchies, might prove to be more a more viable procedure. Baker noted how difficult it is to create tests that are sensitive to academic learning. The solution to this problem, according to Baker, is to design an integrated system in which the starting point is the definition and description of learning objectives or, as the author calls it, task structure. Curriculum design, and ultimately test design, would be based on the defined task structures. Conceptually, the task structure approach to learning uses rules and examples to clearly present a specific set of skills to be learned. Birenbaum and Shaw (1985) also suggested that task structure can be used to develop tests. In this case the skills which a person is expected to learn are modeled and then the test is developed from this model.

TEST DEVELOPMENT

There is evidence that supports the validity of both global and specific measurement for mathematics skills. The choice between the two approaches therefore depends, to a great extent, on a clear understanding of the purposes for which the measuring instruments are being developed.

The Need for Measures of Basic Skills in Mathematics

"A test of `basic skills,'" according to the New Jersey Basic Skills Council (1987), "is a test to determine whether an individual has developed the practical working skills of . . . mathematical literacy needed to take advantage of the learning opportunities that colleges provide" (p. 2). This is equally true in vocational education. As noted above, basic skills in mathematics can be defined as those that the majority of high school graduates would be able to perform successfully after exposure to the typical mathematics curriculum in the educational system of the United States. Instructors in vocational education classes cannot be expected to take time out from substantive topics to teach computational skills which should have been learned at some other point in the student's educational career. Moreover, teaching mathematics is a specialized skill in itself and should be left to instructors skilled in teaching mathematics. Vocational instructors should only be concerned with teaching math as it directly applies to a substantive field.

The need for assessment at the point where the student is just beginning a new course of education is very evident. At a junior college in Texas, a mortality rate of 50% was common in the college algebra course because of an inadequate background in mathematics (Wood, 1980). Of 625 students that took the Texas college placement exam in 1967, 585 were not prepared for college mathematics; consequently, the school began using a placement test to determine if students could be placed directly in college algebra or in a one-semester remedial course. The placement program has since been expanded to the point where, depending on the placement test results, a student can be placed at any of four mathematical ability levels.

The New Jersey Basic Skills Council (1987) reported that 46% of the 1987 class of college freshman lacked proficiency in computational mathematics and 23% of the students were classified as proficient only in some areas of computation. With respect to elementary algebra, 57% lacked proficiency and 29% were proficient in some areas. The figures do not change much if only recent high school graduates are examined. For these recent graduates, 38% lacked proficiency in computations and 25% were partly proficient. Similarly, for elementary algebra, 44% lacked proficiency and 36% were partially proficient. Clearly, a significant number of students would become mortality figures if they attempted regular college mathematics courses or other courses, such as the sciences, where proficiency in computation and elementary algebra are assumed.

Classification of students in the New Jersey Basic Skills assessment program can be made directly from the placement exam scores. The guidelines offered by the Council for Computation follows:

> A scaled score of 164 or below (18 or fewer questions correct out of 30 on the 1987 test) indicates pronounced weaknesses in dealing with certain computational operations and, in particular, with problems involving percentages and decimals. Declining scores indicate progressively greater difficulty with operations involving fractions. Students scoring below 165 on the computation test are included in the category: "Lack Proficiency." The range of scaled scores from 165 to 172 (19 to 24 questions correct) indicates greater familiarity with elementary computation but still shows definite weaknesses. The particular weaknesses of a student can be identified only by examining individual item responses. Students falling in the range of 165 to 172 on the computation test fall in the category: "Appear to be Proficient in Some Areas." Students who achieve a scaled score of at least 174 (25 questions correct) seem to be proficient in the elementary computational skills measured by this test and fall in the "Appear to be Proficient" category. (p. 56-57)

The consequence of providing a "placement service" should be the future success of these students when they return to the normal academic program. Wood (1980) indicates that for 203 students who were identified as needing remedial help, 96.6% improved their scores when they were retook the placement exam at the end of the review. The median score of these students went from 7 to 74. More importantly, however, the mortality rate for the college algebra course dropped from 56% to 28%. The New Jersey Basic Skills Council (1987) reports that the success rate of students who completed the prescribed remedial courses was comparable to non-remedial students, whereas those students who did not complete the remedial review had a success rate of only about one-third of those students who completed remediation.

A vocational assessment program, reported by Benn (1982), demonstrates how specific area tests can be used to place students in appropriate mathematics courses. This project was undertaken to develop a series of program-specific vocational locater tests that would consist of subject-specific questions in three academic disciplines--writing, reading, and mathematics--for use in predicting vocational students' success in their vocational program and to determine what, if any, remedial work they needed to do. The specific skills which a student is presumed to possess as they enter the program were identified. These are skills that the student is expected to have mastered prior to beginning their vocational program and would not be taught in the program. The degree of academic proficiency in mathematics needed for each program was also determined. This information was used to select the specific item bank question used in each of the pilot locater tests. If an examinee did not pass the mathematics portion of the test, the individual items were visually inspected to determine what were the specific deficiencies of the examinee.

The three programs detailed above depict two ways in which assessment of basic mathematics skills has been used with new students. First, the examinee's score on the test is used to decide whether the student is sufficiently prepared in mathematics to proceed with more advanced training in either mathematics or vocational education, or whether the student needs to take remedial courses. Second, for those students deemed mathematically unprepared to learn new material, these programs use crude or inefficient diagnostic measures to determine the level at which new students should begin their remedial education.

Both the New Jersey Basic Skills Program and the testing program at the Texas junior college use a general global score to represent mathematical ability to assign students to the appropriate level of remedial course. If the range of remediation courses is limited, then global estimates of ability may be all that is needed to make a reasonable placement decision; but if the range of remedial education is great (i.e., course in everything from basic addition and subtraction to pre-calculus mathematics), a more accurate measure of ability is needed.

The assessment program described by Benn (1982) is superior to the New Jersey and Texas programs for placement of students in appropriate remedial programs. This program, however, would be difficult to implement on a large scale since it requires visual inspection of the items on each test to determine specific deficiencies. To assess abilities and diagnose deficiencies for a large number of students it would be better to automate the process by using a computer for administration and scoring the test, and also to provide an initial evaluation of which mathematical topics, if any, the student needs to review.

Limitations On The Usefulness Of Global Scores

Even though global scores have been shown to be useful and effective for some situations, there are some major problems that limit their overall usefulness. As Wood (1980) pointed out in a review of mathematics placement procedures used at a Texas junior college, the use of the ACT test score in math for placement decisions decreased the number of failures in college algebra, but the test was, nevertheless, an unreliable measure of algebraic skills since students who had scores above cutoff levels failed college algebra.

The New Jersey Basic Skills Placement Program described above is a good example of the way in which global scores limit the confidence that can be placed in assessment decisions. As indicated, the New Jersey Placement Program uses a single score on the computational test to determine whether the student lacks proficiency, is proficient in some areas, or is possibly proficient in all areas (The New Jersey Basic Skills Council, 1987). The Council feels confident in making classification decisions about students who score below 19 correct answers. They also suggest that if the number of difficult items were increased, they would not need to equivocate about the students who score over 24 by classifying them as "Appear to be Proficient." In point of fact, the same problem does exist for the lower end of the ability scale since more basic skills like addition and subtraction of whole numbers were not tested. More importantly, the global test score cannot determine with precision where students in the middle category, "Appear To Be Proficient In Some Areas," are in relation to the computational hierarchy. For instance, a mid-range score does not provide information on whether the person is having problems with decimals, fractions, or division of decimals and fractions. This limitation is a natural outcome of using a global score with a fixed set of heterogeneous items.

A global score is generally an inadequate measure of specific abilities when the item domain is heterogeneous. Hartke (1978) said that psychometricians generally agree that homogeneity is a desirable characteristic of test items, but questions about what homogeneity means or how it should be measured have not been resolved. To the extent that items measure the same concept or type of performance they are said to be homogeneous (Crocker & Algina, 1986). If a test is composed of vocabulary items, spelling items, math items and history items, it cannot be measuring one type of performance. Even when a test is composed of only math items, more than one type of performance may be being measured (e.g., some items may be strictly computational while others are based on applications).

The degree of item homogeneity of a mathematics test will follow from the definition of the content domain being tested. Tests that measure a variety of mathematical concepts (e.g., adding, multiplying, geometry, algebra) are probably not homogeneous, but tests that measure a particular mathematical concept--single-digit addition, for example--have a high degree of content homogeneity. Between these two extremes is a large gray area where the homogeneity of items may depend on the group taking the exam. For instance, if a test were designed to cover all aspects of addition (e.g., single-digit, multiple-digit, carries, no carries), the test would appear homogeneous for high school students, but might appear to be heterogeneous for third graders.

The description of a skill or trait in terms of the domain being used to test the trait may suggest the use of either heterogeneous or homogeneous items. A heterogeneous domain is necessary if the trait or skill itself is heterogeneous as, for instance, general reading ability. A distinction can be made between response homogeneity, which is based on empirical analysis of response patterns, and conceptual homogeneity, which is based on the conceptual similarity of items. According to Hartke (1978), "The analysis of the conceptual homogeneity of an item population is a logical, judgmental process" (p. 43). Experts make decisions about which sub-population items should be assigned to based on the skills or knowledge examinees must have to correctly respond to the items. These two conceptualizations of domain homogeneity coincide when the conceptual definition is so narrow that the domain items would tend to have roughly equivalent item characteristics (e.g., two-digit multiplication problems). In this case, the items are completely interchangeable. Up to a point, even when the domain definition is loosened (e.g., all multiplication problems involving whole numbers), as long as the items are conceptually similar, items may be treated as interchangeable for testing purposes.

One advantage to using highly homogeneous tests is that the test user can have greater certainty in the interpretations that are made of a test score (Anastasi, 1976). If the test is heterogeneous, any particular score could represent many different patterns of ability. On the other hand, with a highly homogeneous set of items a particular score could be interpreted as the proportion of the content domain that the person knows or, if the items are arranged in increasing order of difficulty, the score can be treated as a rough estimate of the person's ability within the narrow domain of items. Another advantage of having highly homogeneous items is that fewer items are needed for measurement at any given degree of accuracy (Shoemaker, 1975). The number of items needed to measure a domain at a given level of accuracy increases as the item heterogeneity increases or the variance of item difficulties increases. In practical terms, a person's test behavior would be expected to be consistent within homogeneous domains or subdomains, and variable across heterogeneous domains (Hively, Patterson, & Page, 1968).

Using a global score to represent an examinee's achievement across several mathematical domains, instead of scores on the separate domains, would create two problems. One difficulty would be that no precise interpretation could be given to the score because of the high level of item heterogeneity. In particular, no information on the pattern of performance on separate, homogeneous domains would be available. This problem is evident in the New Jersey Placement Program. Another problem would be the increase in the number of items needed to generate scores with sufficient accuracy.

Another potentially serious problem which could occur when using global scores for classification is that the scores may be biased against a subclass of examinees. As a consequence of the American education system, where local school districts have authority to set curriculum, the curriculum varies between school districts. Because of these differences, test publishers must expend resources to assure that tests are adequate measures of educational objectives common to a large number of schools (Mehrens & Phillips, 1986), but even for large commercial test publishers there is no guarantee that there will be a perfect match between any particular curriculum and an achievement test. Airasian and Mandaus (1983) suggest that many achievement tests, because of their widespread use, are designed to assess only those skills that all schools have used. In effect, tests are designed to assess the lowest common denominator of skills. Standardized tests, according to Airasian and Mandaus, used in this way are not sensitive to differences between schools or programs because the tests do not adequately reflect specific content and objectives of particular schools and programs. Haladyna and Roid (1981) refer to this as "instructional sensitivity" and define it "as the tendency for an item to vary in difficulty as a function of instruction" (p. 40).

Coombs (1941), in a factorial study of numerical ability, hypothesized that this ability might be related to "quick recollection and manipulation of well-established associations" (p. 164). Coombs pointed out that these associations come about through education; therefore, a person whose education did not include training specific to developing quick recollection of these associations might have lower scores on this ability than if their education did include the relevant training. Mehrens and Phillips (1986) refer to this as "content tested but not taught." For individuals and groups who did not receive instruction on the content represented by items, these items are non-representative--and, hence, invalid and biased, according to Schmidt (1983).

Research by Phillips and Mehrens (1987) suggests that the potential bias problem may not be as great as feared; nevertheless, this problem can be expected to occur at least for some individuals such as foreign students or even those who come from an atypical school district. While bias may have a negative impact on a particular individual if global scores are used (this is test bias for which the test developer and users may be held accountable), there is no "bias" in diagnostic testing because the source of bias is not within the test. This is because in diagnostic assessment the only thing the evaluator is interested in is whether the student has the prerequisite skills necessary to complete a particular course of study. If they lack the skills, for whatever reason, they need to be offered remedial education.

Alternatives to Global Measurement

Ruthven (1987) summed up the natural tension which arises from using global assessment to evaluate mathematical learning very well when he said:

> It seems, then, that the view of mathematics learning as an ordered progression through a hierarchy of knowledge and skill, medicated by a stable cognitive capability of the individual pupil . . . is defensible only as a gross, general, global model. It appears to be incapable of capturing the finer, more individual and more local aspects of mathematics learning, and thus to be of limited value in describing and understanding the particular cognitive capabilities of individual pupils in order to plan, promote and evaluate their learning. (p. 247)

Clearly, global achievement scores, which are based on a heterogeneous content domain, do not provide very much information about the specific skills and abilities of the examinees. Even when the domain definition can be made clear and concise, the assessment of mathematical ability by a single global achievement score is inadequate. Only when "a single thread of relationships ties together most of the categories in the universe" would we be able to deduce from a total score which skills a person has (Hively, Patterson, & Page, 1968). But when the subject matter has a complex structure such as in mathematics where there are independent, hierarchically structured areas, a single score cannot explain the various patterns of scores that add up to the total score. On the other hand, narrowly defined domains are very useful when the test user is interested in whether particular skills have been learned. (Linn, 1980). The key elements for designing a program to assess the mathematical ability of students entering a vocational education program is to accurately define the domain structure and provide an assessment battery which efficiently determines the mathematical ability of the examinee within the total domain structure.

Conventional paper-and-pencil tests, such as the Texas junior college placement exam or the New Jersey test of basic skills, which are designed to cover a broad ability band using a fixed item format, use valuable testing time inefficiently. When an item is much too easy or far too difficult for the individual at a specific ability level, little useful information is obtained about that individual's ability.

One assessment methodology which may be adaptable to use as an assessment battery design comes from adaptive or sequential testing research. Adaptive testing is a flexible system of testing which allows for different sets of items to be given to individual examinees depending on the characteristics of the individual. Weiss (1985) provides a basic introduction to adaptive testing.

Efficient testing may be important when assessing ability in a single content domain; efficient testing procedures are critically important when assessment is being done across several content domains, such as with a mathematical ability test battery. Reise and Weiss (1990) evaluate efficient computer-administered adaptive and sequential approaches to measuring mastery of specific skills. These approaches would be particularly useful in mathematics skills assessment where subcontent domains can be narrowly defined.

Adaptive or sequential testing approaches can also take advantage of the hierarchies extant in mathematics achievement, using approaches such as proposed by Brown and Weiss (1977). The authors compared the effects of using a combination of inter-content and intra-content branching strategy for a five content area biology test against results from a conventional test for the same content areas. The unique aspect of this strategy was that the estimated ability of an examinee at the end of one subtest was used, in combination with a measure of the degree of redundancy between the first test and the next test to be administered, to determine the starting point in subsequent subtests. Thus, after the first test examinees would have differential starting points in all subsequent tests. As anticipated from prior research, the average reduction in test length was 49.3%--a substantial savings in testing time. More importantly, the data demonstrated that the minimum number of items administered under the adaptive strategy decreased as latter subtests were administered. The authors attribute this savings to the increased use of prior information.

If assessment is being conducted for reasons such as assessment of mathematical ability for diagnostic and placement purposes, it may not be necessary to assess the examinee's performance in all content areas when the separate content domains are hierarchically structured. Spineti and Hambleton (1977) suggested that adaptive testing methods could substantially reduce testing time when the domain has a hierarchical structure of learning objectives. In their simulation study, Spineti and Hambleton investigated the effects on the quality of decisions and testing time of varying several test relevant factors. Test length for each objective was set at 1, 2, 3, 4, and 5 items; the mastery cutting score was set at either 3, 4, or 5; and four different starting points for each of two hierarchies were used. One hierarchy used in the study was initially developed by Gagne (1965) to represent the learning structure for hydrolysis of salts. The second hierarchy came from Ferguson (1969) and concerned addition and subtraction. One interesting feature of this study was the non-computerized routing strategy used. The routing strategy had two basic rules: (1) wherever in the hierarchy the examinee started, they moved sequentially up or down the hierarchy depending on whether they mastered the objective; and (2) if an examinee was successful at a mastery level they were given credit for all objectives below that point in the hierarchy--if the person was unsuccessful, they were assumed to be unable to be successful at any higher level in the hierarchy.

Results from the Spineti and Hambleton study for the Gagne hierarchy indicate that there was a 59.2% average reduction in testing time with slightly fewer classification errors. Results from the analyses of Ferguson's hierarchy showed a similar pattern. For the adaptive test, on the average, only 8.43 o: the 18 objectives were tested with essentially equivalent errors to the conventional test. Also, results indicate that starting in the middle of the hierarchy produced the greatest savings in test time. On the average, two less objectives needed to be tested if testing were started in the middle of the hierarchy.

CONCLUSIONS

In order to develop a diagnostic test battery for measuring mathematical ability many separate issues must be resolved. Initially, the domain structure must be clearly defined so that test items can be written to reflect its elements. The conceptual literature concerning the domain of mathematics suggests that there are branches of mathematics such as algebra and topology which share few, if any, common elements, but mathematical concepts within a branch of mathematics such as algebra are structured hierarchically. The conceptual definitions of mathematical structure are interesting and may be useful for some purposes, but they are not specific enough for test development purposes. They do, however, suggest that an adequate test for measuring basic skills in mathematics should take into account the hierarchies of skills that exist within defined areas of mathematics.

Describing the domain of mathematics solely in terms of its mathematical properties is an important first step, but the development of a test battery for measuring mathematical skills also requires an understanding of the developmental processes of the persons to be examined. Basically, developmental theorists suggest that children progress through an ordered sequence of levels of cognitive understanding of mathematics. In the early stages a person is mainly concerned with learning very specific computational algorithms; it is only much latter in the developmental process that a person can perform higher-order mathematical operations. Observations about ability that are made from a developmental point of view, however, tend to be a global evaluation of the person's ability and therefore such evaluations are of limited usefulness for diagnostic purposes. Similar to the conceptual definitions of mathematics, however, the developmental theorists also emphasize the hierarchical structure of the domain.

The work of researchers and authors such as Bloom and Wilson in curriculum development provides the necessary structure for relating the domain of mathematics to the developmental processes of the target population. Wilson (1971) accomplishes this by creating a two-dimensional matrix where mathematical content is represented on one dimension and the cognitive complexity of the problem on the other dimension. The cross-classification of both dimensions into combined categories leads to a useful description of the individual cells of the matrix. The matrix cells, which represent behavior objectives, provide all the information necessary to generate test items to assess a person's competency with respect to a particular cell. However, for purposes of measuring the basic math skills of high school graduates, Wilson's taxonomy would have to be extended to a greater range of content than that covered by the seventh through ninth grade curriculum on which it was based.

The nature of mathematical ability has been empirically investigated in two major ways. First, a review of factor analytic studies of mathematical ability indicates that there is a general ability factor for mathematics, as well as separate factors representing different content domains in mathematics. The general factor of mathematical ability is strongest when the test battery is made very heterogeneous by including other kinds of tests such as verbal ability, whereas group factors become stronger when the tests include only mathematics items. The factors, both general and specific, retained for analysis together often account for less than 50% of the total variance, because many factors are defined by one or two items and so do not account for much of the common variance. As a consequence, they have been typically eliminated from the analysis. The fact that many studies provide evidence that a few specific factors can account for most of the common variance may be misleading since in most studies 50% or more of the matrix variance is still not accounted for by the extracted factors.

A second major line of empirical research is typified by the work of Gagne and others who have shown that clusters of mathematical concepts can be interrelated within a hierarchical structure. Basically, two tasks are hierarchically related if one task must be learned prior to learning another task. A hierarchical structure can be described a priori using the method developed by Gagne, by more sophisticated methods such as through task analysis, or empirically. Although the evidence is not unequivocal, there is sufficient evidence from numerous studies to support the conclusion that mathematical concepts or processes can be hierarchically related. Hierarchical structures have been shown to exist for computational problems, working with algebraic equations, and graph interpretation skills. One limitation of these studies, however, is that they concerned circumscribed areas of mathematics. Research needs to be done to determine if hierarchies also exist across mathematical content areas, as well as research in mapping out hierarchical relationships within narrowly defined areas of mathematics.

It is quite evident that there is a real need for accurate, efficient diagnostic testing of mathematical abilities. As the experience of the New Jersey Basic Skills Program (1987) and the Texas junior college program (Wood, 1980) indicates, a significant proportion of students who graduated from high school needed to take remedial courses in mathematics. Programs such as these, however, make only relatively crude placement decision about students who were diagnosed as deficient in mathematical skills because they rely on a single global score to represent mathematics achievement rather than focusing on an individual's pattern of mathematical abilities. Measuring basic skills in mathematics using global scores may be useful for distinguishing between persons who are either competent or not competent in these skills; however, they cannot be used with any accuracy to prescribe specific remediation courses for the examinee. Furthermore, in certain situations, such as inadequate instruction, global scores may be biased estimates of a examinee's mathematical ability. Diagnostic assessment, which is only concerned with determining whether a person has mastered a very specific skill, should be less susceptible to bias related problems.

Finally, complete diagnostic testing of mathematical ability would be very time-consuming if paper-and-pencil tests are used. Even if only three questions per skill were used, a battery developed from Wilson's (1971) behavior objective matrix would have several hundred items. Research done by Brown and Weiss (1977) and Spineti and Hambleton (1977) has shown that computerized testing methods can, on the average, significantly reduce the number of items which must given to an examinee with no loss of measurement accuracy. A further reduction in testing time is possible if a computerized adaptive testing system can utilize the hierarchical structure of mathematics by not testing at all levels of the hierarchy. In addition to providing maximal diagnostic capability in a minimal amount of student testing time, computerized administration of a mathematics basic skills battery would also allow immediate test scoring and reporting, thereby providing test data to the student and instructor in a time frame and form that would maximize its use in the educational process.

REFERENCES

Airasian, P. W. & Bart, W. M. (1975). Validating a priori instruc tional hierarchies. *Journal of Educational Measurement, 12*, 163-173.

Airasian, P. W. & Mandaus G. F. (1983). Linking testing and instruction: Policy issues. *Journal of Educational Measurement, 20*, 103-118.

Anastasi, A. (1976). *Psychological testing* (5th ed.). New York: Macmillian.

Baker, E. L., & Herman, J. L. (1983). Task structure design: Beyond linkage. *Journal of Educational Measurement, 20*, 149-164.

Bart, W. M. & Krus, D. J. (1973). An ordering-theoretic method to determine hierarchies among items. *Educational and Psychological Measurement, 33*, 291-300.

Benn, R. J. (1982). *Development of a program specific locater test. Final Report.* Tocoma, WA: Fort Steilacoom Community College. (ERIC Document Reproduction Service No. ED229595)

Birenbaum, M., & Shaw, D. J. (1985). Task specifications chart: A key to a better understanding of test results. *Journal of Educational Measurement, 22*, 219-230.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning.* New York: McGraw-Hill.

Brown, J. M., & Weiss, D. J. (1977). *An adaptive testing strategy for achievement test batteries* (Research Report 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, October 1977.

Bruner, J. S. (1960). On learning mathematics. *Mathematics Teacher, 53*, 610-619.

Coombs, C. H. (1941). A factorial study of number ability. *Psychometrika, 6*, 161-189.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt.

Dienes, Z. P. (1960). *Building up mathematics.* London: Anchor.

Dienes, Z. P. (1973). *The six stages in the process of learning mathematics.* NFER (English translation of 1970 French original).

Ferguson, R. L. (1969). The development of a computer-assisted branched test for a program of individually prescribed instruction. Unpublished doctoral dissertation, University of Pittsburgh.

Freemont, H. (1969). *How to teach mathematics in secondary schools.* Philadelphia: Saunders.

Fruchter, B. (1952). Orthogonal and oblique solutions of a battery of aptitude, achievement and background variables. *Educational and Psychological Measurement, 12,* 20-38.

Furneaux, W. D. & Rees, R. (1978). The structure of mathematical ability. *British Journal of Psychology, 69,* 507-512.

Gagne, R. M. (1962). The acquisition of knowledge. *Psychological Review, 69,* 355-365.

Gagne, R. M. (1965). *The conditions of learning.* New York: Holt, Rinehart, and Winston.

Gagne, R. M. & Paradise, N. E. (1961). Abilities and learning sets in knowledge acquisition. *Psychological Monographs: General and Applied, 75,* 1-23.

Gagne, R. M., Mayor, J. R., Garstens, H. L., & Paradise, N. E. (1962). Factors in acquiring knowledge of a mathematical task. *Psychological Monographs: General and Applied, 76,* 1-26.

Garden, R. A. & Robitaille, D. F. (1989). Test development, scoring and interpretation. In D. F. Robitaille & R. A. Garden (Eds.), *The IEA study of mathematics II: Contexts and outcomes of school mathematics,* (pp. 84-101). Oxford: Pergamon Press.

Green, R. F., Guilford, J. P., Christensen, P. R. (1953). A factor-analytic study of reasoning abilities. *Psychometrika, 18,* 135-160.

Griffiths, H. B. & Howson, A. G. (1974). *Mathematics: Society and curricula.* London: William Clowes and Sons.

Guion, R. M. (1977). Content Validity--the source of my discontent. *Applied Psychological Measurement, 1,* 1-10.

Haertel, E. & Calfee, R. (1983). School achievement: Thinking about what to test. *Journal of Educational Measurement, 20,* 119-132.

Haladyna, T. & Roid, G. (1981) The role of instructional sensitivity in the empirical review of criterion-referenced test items. *Journal of Educational Measurement,* 18, 39-53.

Hambleton R. K. (1980). Contributions to criterion-referenced testing technology: An introduction. *Applied Psychological Measurement, 4,* 421-424.

Harman, H. H. (1976). *Modern factor analysis (3rd ed.).* Chicago: University of Chicago Press.

Hart, K. (1981). Hierarchies in mathematics education. *Educational Studies in Mathematics, 12,* 205-218.

Hartke, A. R. (1978). The use of latent partition analysis to identify homogeneity of an item population. *Journal of Educational Measurement, 15,* 43-47.

Hieronymous, A. N. (1972). Evaluation and reading: Perspective '72. *The Reading Teacher, 26,* 264-267.

Hively, W., II, Patterson, H. L., & Page, S. H. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement, 5,* 275-290.

Howson, G., Keitel, C. & Kilpatrick, J. (1981). *Curriculum development in mathematics.* Cambridge: Cambridge University Press.

Kline, A. E. (1981). Redundancy in the Iowa Tests of Basic Skills. *Educational and Psychological Measurement, 41,* 537-544.

Kline, M. (1962) *Mathematics: A cultural approach.* MA: Addison-Wesley.

Kolb, J. R. (1967). Effects of relating mathematics to science instruction on the acquisition of quantitative science behaviors. *Journal of Research in Science Teaching, 5,* 174-182.

Kolen, M. J., & Jarjoura, D. (1984). Item profile analysis for tests developed according to a table of specifications. *Applied Psychological Measurement, 8,* 321-331.

Kothe, G., & Ballier, F. (1974) The changing structure of modern mathematics. In H. Behnke, F. Bachmann, K. Fladt & W. Suss (eds), *Fundamentals of Mathematics* (pp. 505-528). Cambridge, MA: MIT Press.

Lawrence, I. M. & Dorans, N. J. (1987, April). *An assessment of the dimensionality of SAT-Mathematical.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, D.C.

Linke, R. D. (1975). Replicative studies in hierarchical learning of graphical interpretation skills. *British Journal of Educational Psychology, 45,* 39-46.

Linn, R. L. (1980). Issues of validity for criterion-referenced measures. *Applied Psychological Measurement, 4,* 547-561.

Martin, D. J., & Dunbar, S. B. (1985). Hierarchical factoring in a standardized achievement battery. *Educational and Psychological Measurement, 45,* 343-351.

Mathematics activities, level nine. Teachers Guide. (1982). Philadelphia, PA: Office of Curriculum and Instruction. (ERIC Document Reproduction Service No. ED 239 845)

Mehrens, W. A., Phillips, S. E. (1986). Detecting impacts of curricular differences in achievement test data. *Journal of Educational Measurement, 23,* 185-196.

New Jersey Basic Skills Council. (1987). *New Jersey College Basic Skills Placement Testing. Fall 1987. Entering Freshmen.* Trenton, NJ: New Jersey State Department of Higher Education, New Jersey Basic Skills Council. (ERIC Document Reproduction Service No. ED 292 375)

Northrop, L. C. (1977). *The definition and measurement of numerical ability* (CSC report no. PRDC-TM-77-9). Washington, D.C: Civil Service Commission, Personnel Measurement Research and Development Center. (ERIC Document Reproduction Service No. ED 150 179)

Phillips, S. E. & Mehrens W. A. (1987). Curricular differences and unidimensionality of achievement test data: An exploratory analysis. *Journal of Educational Measurement, 24,* 1-16.

Piaget, J. (1952). *The child's conception of number.* New York: Norton.

Powers, D. E., Swinton, Spencer S., & Carlson, A. B. (1977). *A factor analytic study of the GRE aptitude test* (GRE Board Professional Report GREB No. 75-110). Princeton: Educational Testing Service. (ERIC Document Reproduction Service No. Ed 163 081).

Reckase, M. D., Davey, T. & Ackerman, T. (1989, March). *Similarity of the multidimensional space defined by parallel forms of a mathematics test.* Paper presented at the Annual Meeting of American Educational Research Association, San Francisco, CA.

Rees, M. (1962). The nature of mathematics. *The Mathematics Teacher, 55,* 434-440.

Reise S. P. & Weiss, D. J. (1990). *Methods for computerized mastery testing* (Unpublished manuscript). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Resnick, L. B. (1973). Hierarchies in children's learning: A symposium. *Instructional Science, 2,* 311-362.

Resnick, L. B., & Ford, W. W. (1981). *The psychology of mathematics for instruction.* New Jersey: Lawrence Erlbaum Associates.

Robitaille, D. & Dirks, M. (1982). Models for the mathematics curriculum. *For the Learning of Mathematics, 2,* 3-21.

Robitaille, D. F., Sherrill, J. M. & O'Shea, T. J. (1980). *Mathematics Achievement Test Project. Technical Manual.* Victoria, British Columbia: British Columbia Department of Education. (ERIC Document Reproduction Service No. ED 187531)

Russell, B. (1919). *Introduction to mathematical philosophy.* London: Allen and Unwin.

Ruthven, K. (1987). Ability stereotyping in mathematics. *Educational Studies in Mathematics, 18,* 243-253.

Schmidt, W. H. (1983). Content biases in achievement tests. *Journal of Educational Measurement, 20,* 165-178.

Shepard, L. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4,* 447-468.

Shermis, M. D. (1988, April). *The use of IRT to investigate the linear/hierarchical nature of a college mathematics curriculum.* Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Shoemaker, D. M. (1975). Toward a framework for achievement testing. *Review of Educational Research, 45,* 127-147.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology, 15,* 201-293.

Spineti, J. P. and Hambleton, R. K. (1977). A computer simulation study of tailored testing strategies for objective-based instructional programs. *Educational and Psychological Measurement, 37,* 139-158.

Spitznagel, E. L., Jr. (1971). *Selected topics in mathematics.* New York: Holt, Rinehart and Winston.

Steen, L. A. (1978) Mathematics today. In Lynn A. Steen (ed.), *Mathematics Today: Twelve Informal Essays* (pp. 1-12). New York: Springer-Verlag.

Stein, Sherman K. (1963). *Mathematics: The man-made universe.* San Francisco: Freeman.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs.* No. 1. Chicago: University of Chicago Press.

Very, P. S. (1967). Differential factor structures in mathematical ability. *Genetic Psychology Monographs, 75,* 169-207.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53,* 774-789.

Weiss, D. J. & Betz, N. E. (1973). *Ability measurement: conventional or adaptive?* (Research report 73-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1973. (NTIS No. AD 757788)

White, R. T. (1973). Research into learning hierarchies. *Review of Educational Research, 43,* 361-375.

Wilson, J. W. (1971). Evaluation of learning in secondary school mathematics. In B. S. Bloom, J. T. Hastings, & G. F. Madaus (Eds.), *Handbook on formative and summative evaluation of student learning* (pp. 643-696). New York: McGraw-Hill.

Wood, J. P. (1980). Mathematics placement testing. *New Directions for Community Colleges, 31,* 59-64.

Wrigley, J. (1958). The factorial nature of ability in elementary mathematics. *British Journal of Educational Psychology, 28,* 61-68.