UNIVERSITY OF ALBERTA

EVALUATING BLOOM'S COGNITIVE LEVELS IN THE TABLE OF SPECIFICATIONS FOR A GRADE 6 MATHEMATICS ACHIEVEMENT TEST

BY

©

MARK J. GIERL

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree MASTER OF EDUCATION.

in

School Psychology

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

Edmonton, Alberta, Canada

FALL, 1993

ISBN 0-315-88238-7

UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR: Mark J. Gierl

TITLE OF THESIS: Evaluating Bloom's Cognitive Levels in the Table of Specifications for a Grade 6 Mathematics Achievement Test

DEGREE: Master of Education

YEAR THIS DEGREE GRANTED: Fall, 1993

11025 - 84 Ave

Edmonton, AB

Canada T6G-0V6

May, 1993

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommended to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled EVALUATING BLOOM'S COGNITIVE LEVELS IN THE TABLE OF SPECIFICATIONS FOR A GRADE 6 MATHEMATICS ACHIEVEMENT TEST submitted by MARK J. GIERL in partial fulfillment of the requirements for the degree of MASTER OF EDUCATION.

Thomas O. Maguire

W. Todd Rogers

Jeffrey Bisanz

Dwight Harley

May 26, 1993

Abstract

The purpose of this study was to determine whether the Taxonomy of Educational Objectives: Cognitive Domain (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956) provides an accurate model to guide item writers for anticipating the cognitive processes used by Grade 6 students to solve items on a large-scale achievement test in mathematics. The validity of the cognition section in the table of specifications for an achievement test depends on two assumptions that were tested in this study. The first is that students use the cognitive processes anticipated by item writers in the same proportions as outlined in the table. To investigate this assumption 30 Grade 7 students (16 males, 14 females) were asked to think aloud as they solved problems on a mathematics achievement test that contained 18 items previously classified and administered by Alberta Education. Students' cognitive processes were classified using a coding system based on Bloom's taxonomy. The overall match between the responses expected by item writers and observed from students was 53.7%. A comparison of response frequencies indicated that students used the cognitive processes described in Bloom's taxonomy, but not in the same proportions as anticipated by item writers. The second assumption tested in this study was that high and low math achievers use the cognitive processes anticipated by item writers in the same proportions as outlined in the table of specifications. To investigate this assumption the same students

were rank-ordered into two achievement categories according to their teacher-assigned math grades. Separate analyses using the response frequencies of the high and low math achievers indicated that the two groups tended to use similar cognitive processes to solve items but not in the same proportions as anticipated by item writers. Agreements between the expected and observed cognitive processes were assessed with an analysis of variance. A three-way interaction of achievement group, cognitive level, and content area was found. Post-hoc comparisons revealed that the match between expected and observed responses was highest for comprehension items and that item writers were more accurate at anticipating the cognitive processes used by high math achievers. When student protocols on select items were examined, many different problem-solving strategies were identified demonstrating that the levels in Bloom's taxonomy, as used in test construction, conceal response variability. Overall, the results of this study indicated that Bloom's taxonomy does not provide an accurate model to guide item writers for anticipating the cognitive processes used by students.

Acknowledgement

I would like to thank the members of my thesis committee for their expertise and guidance throughout this project, and throughout my Master of Education training: Tom Maguire for his careful yet unconfining supervision that has allowed me to become a more independent researcher and thinker; Todd Rogers and Dwight Harley for their technical expertise and insightful critiques; and Jeff Bisanz for his presence and participation that helped me ensure a little psychology always accompanied a good statistical analysis.

I extend my gratitude to the students and staff at Louis St. Laurent school in Edmonton, Alberta, for their cooperation, as well as to Dennis Belyk at Alberta Education for helping me obtain materials and information related to the provincial achievement testing program.

I also appreciate the support and patience provided, once again, by Jeannie Dominey and my parents.

TABLE OF CONTENTS

I. INTRODUCTION.. 1

A. Purpose of the Study........................ 4

B. A Cognitive Critique of Psychometric Models... 7

C. A Summary and Critique of the Cognitive Domain in the Taxonomy of Educational Objectives..... 10

D. Preview of the Present Study.................. 18

II. METHOD.. 19

A. Subjects...................................... 19

B. Materials..................................... 20

C. Procedure..................................... 25

III. RESULTS.. 25

A. Mathematics Achievement Subscale.............. 25

B. Inter-rater Agreement......................... 26

C. The Cognitive Processes Used by Students Overall....................................... 27

D. The Cognitive Processes Used by High and Low Achievers..................................... 29

E. Evaluating the Knowledge and Application Items Coded as Comprehension................. 39

IV. SUMMARY.. 45

V. GENERAL DISCUSSION................................ 48

REFERENCES.. 53

APPENDIX.. 59

LIST OF TABLES

TABLE 1. The Six Major Levels in the Cognitive Domain of the Taxonomy of Educational Objectives...... 13

TABLE 2. The Mathematics Achievement Subtest Contingency Table for the Expected (Rows) and Observed (Columns) Cognitive Responses as a Function of Achievement Group and Content Area.. 28

TABLE 3. Mean Agreement Scores For Expected and Observed Responses on Knowledge, Comprehension, and Application as a Function of Achievement Group and Sex................................ 32

TABLE 4. Summary Table for the 2 (Achievement) x 2 (Sex) x 3 (Cognitive Level) x 2 (Content Area) Analysis of Variance With Repeated Measures on the Last Two Variables............. 34

TABLE 5. Summary of Student Protocol Responses as a Function of the Cognitive Process Expected by Alberta Education and Observed from High and Low Math Achievers........................... 40

LIST OF FIGURES

FIGURE 1. The Table of Specifications used by Alberta Education to assess Grade 6 Mathematics Achievement in 1991.......................... 5

FIGURE 2. The Table of Specifications for the Mathematics Achievement Subscale (MAS)........ 24

Evaluating Bloom's Cognitive Levels in the Table of Specifications for a Grade 6 Mathematics Achievement Test

Large-scale achievement testing is an important feature of education in North America. Achievement tests include school leaving exams, minimum competency tests, and diagnostic assessments. These tests are developed by a variety of agencies, ranging from private testing companies to governmental branches, who evaluate whether students have attained the goals of schooling. The goals of schooling are numerous, and many of these goals include changes in student's cognitive skills. For example, students are expected to comprehend principles in mathematics and evaluate arguments in social studies.

Assessing cognition using achievement tests is difficult. Test developers try to overcome this difficulty by considering the curricular, cognitive, and predictive features of the achievement test (Millman & Greene, 1989). Often, however, the emphasis during test construction is on curricular features such as content coverage (Emmerich, 1989) and predictive features such as student classification (Embretson, 1985). Cognitive features, such as strategy selection and higher-order thinking, are often poorly evaluated because item writers are not trained to identify the cognitive processes required to solve test items. In most cases, item writers are content specialists working from test specifications that have no formal relation to contemporary psychological theory (Embretson, 1985; Snow & Peterson, 1985).

Currently, the most widely used model for identifying the cognitive processes used by examinees to solve test items is the Taxonomy of Educational Objectives: Cognitive Domain created by Bloom, Englehart, Furst, Hill, and Krathwohl (1956). Bloom's taxonomy is popular because it focuses on a list of thinking skills commonly used in education and it provides a standard vocabulary for classifying learning outcomes (Gronlund, 1982, 1991; Osterlind, 1989; Smith, 1984; Roid & Haladyna, 1982). However, Bloom's taxonomy is a model of *cognitive intentions* and may not be an accurate description of the cognitive processes that students actually use when responding to achievement test items. In short, little is known about the validity of using Bloom's taxonomy for classifying cognitive processes, and about the processes and strategies students actually use to solve items on achievement tests.

The impact of Bloom's taxonomy in test design is most apparent in the table of specifications. The purpose of the table is to outline the achievement domain and provide a guideline for obtaining a representative sample of test items. Although the structure of the table can vary, one of the most common procedures is to create a two-dimensional matrix (Osterlind, 1989; Ebel & Frisbie, 1986; Gronlund, 1991; Smith, 1984). One dimension of the matrix specifies the cognitive objectives: How students are expected to think about and respond to a test item. Although many different learning outcomes can be represented in this dimension, such as attitudes and psychomotor skills, cognitive skills are commonly specified for achievement tests. As previously mentioned, Bloom's taxonomy is one of the most widely used classification systems for labeling cognitive processes. The Taxonomy of Educational Objectives provides a systematic outline of six different levels of thinking that were proposed as goals of classroom instruction (Bloom et al., 1956). The taxonomy begins with the simplest level - knowledge (i.e., recall of specific information) and ends with the most complex level - evaluation (i.e., ability to judge the value of materials and methods for given purposes). Items for each cognitive level in the table of specifications are created by writers who try to anticipate the cognitive processes examinees will use to answer the questions correctly (Millman & Greene, 1989).

The second dimension of the matrix specifies the content objectives: The topical areas covered during instruction. The content covered on a test should reflect the emphasis it was given during instruction. Together, the cognitive level and content area dimensions form cells in the table of specifications. Each cell contains a number of test items. The degree to which the test represents the table of specifications will determine how well the test measures what it was designed to assess. The table is also used for interpretive purposes as the score for a cluster of items in a cell can help the user evaluate differential test performance (i.e., how well students or groups of students perform in relation to different cells that are assumed to measure different cognitive processes in different content areas).

To illustrate how Bloom's taxonomy is used in test construction, the table of specifications used by Alberta Education in 1991 to assess achievement in Grade 6 mathematics is provided in Figure 1. The cognitive dimension is on the top of the table and the content dimension (called Curricular Elements by Alberta Education) is on the left side. The numbers in each cell represent a test item measuring one cognitive skill in one content area. Only the first three levels in Bloom's taxonomy are measured with this achievement test: knowledge, comprehension, and application. Application skills (43.6% of cognitive coverage) receive the most emphasis on this exam. Six content areas are measured with this test: numeration, operations and properties, measurement, geometry, graphing, and problem-solving strategies. Numeration, operations and properties, and measurement (72.7% of the content coverage) receive the most emphasis on this exam.

<u>Purpose of the Study</u>

The purpose of this study was to determine whether Bloom's taxonomy provides an accurate model to guide item writers for

Figure 1. The table of specifications used by Alberta Education to assess Grade 6 mathematics achievement in 1991.

Curricular Elements	Component		Total	Cognitive Level			Total
	Problem Solving	Subject Matter		Knowledge	Compre-hension	Application and Problem Solving	
Numeration	4, 53, 54, 55	1, 2, 3, 5, 6, 14, 15, 38, 39	13 [23.6%]	1, 2, 14, 38, 39	3, 5, 6, 15	4, 53, 54, 55	13 [23.6%]
Operations and Properties	20, 27, 36, 41, 47, 48, 50	29, 37, 40, 42, 43, 44, 45, 46, 49	16 [29.1%]	42	29, 36, 37 ,40, 43, 44, 45, 46, 49	20, 27, 41, 47, 48, 50	16 [29.1%]
Measurement	16, 28, 35, 51, 52	7, 8, 11, 12, 13, 17	11 [20.0%]	7, 12, 13, 17	8, 11	16, 28, 35, 51, 52	11 [20.0%]
Geometry	32	9, 30, 31, 33, 34	6 [10.9%]	9, 30	31, 34	32, 33	6 [10.9%]
Graphing	21, 22, 23, 24, 25	26	6 [10.9%]	—	24, 26	21, 22, 23, 25	6 [10.9%]
Problem-Solving Strategies	10, 18, 19	—	3 [5.5%]	—	—	10, 18, 19	3 [5.5%]
Total	25 [45.5%]	30 [55.5%]	55 [100%]	12 [21.8%]	19 [34.5%]	24 [43.6%]	55 [100%]

anticipating the cognitive processes used by elementary school students to solve items on a large-scale achievement test in mathematics. To achieve this goal, two main issues were addressed. First, items for each cognitive level in the table of specifications are developed by writers who infer the cognitive processes required by the examinee to answer the questions correctly. The validity of this technique was assessed by addressing the question: Do students use the cognitive processes identified by item writers in the same proportions as outlined in the table of specifications? For example, if three items designed to measure knowledge processes are solved by 30 students, are 90 knowledge responses identified? Second, the table of specifications does not differentiate examinees. Consequently, high and low mathematics achievers are assumed to use the same cognitive processes to solve test items. The validity of this assumption has not been examined and an important question remained: Do high and low mathematics achievers use the cognitive processes identified by item writers in the same proportions as outlined in the table of specifications?

A critique of the models commonly used in educational measurement from a cognitive perspective provides a starting point for the present study, and is followed by a summary and critique of the cognitive domain in the <u>Taxonomy of Educational Objectives</u>.

A Cognitive Critique of Psychometric Models

An introductory textbook approach to cognitive psychology provides chapters on basic information processing such as perception, memory, language, and thinking (Best, 1989). Cognitive psychologists also study academic skills such as school learning in reading, writing, and mathematics. Research on complex educational tasks is now becoming an important topic in cognitive psychology because it provides a deeper understanding of how mental processes interact with content areas to influence learning (Snow & Lohman, 1989), and it reflects a societal demand as students are expected to develop more complex thinking skills in school (Resnick, 1987). The same arguments are motivating cognitive psychologists to study educational tests: Tests are complex educational tasks where cognitive processes interact with content areas to provide a measure of learning, and tests are valued by many members of society.

Cognitive researchers have identified three weaknesses in the psychology of psychometric models. First, the psychometric models used in classical test theory, item response theory, and generalizability theory offer little psychological justification for how examinees answer items. Rather, the models are often justified by how well they describe the empirical results. For example, Snow and Lohman (1989) quote Lord (1980) who states that no psychological theory guides the three-parameter item response model, instead "the model must be justified on the basis of the results obtained" (p. 14). Thus, psychometric models are often judged on utilitarian and empirical rather than theoretical and psychological grounds.

Second, psychometric models often make simplistic and unrealistic assumptions about the psychology of test performance. Snow and Lohman (1989) point to classical test theory and generalizability theory as examples. For the models used in these two theories, test-item errors are assumed to be uncorrelated. This implies that examinees do not learn during a test, do not react to their successes and failures on previous items during a test, and do not change their behaviors as the test progresses. However, research fails to support this assumption. Examinee experience appears to be a key variable in determining what attributes a test measures. For example, when the test performance of novices and experts is compared, the two groups show different problem-solving behaviors (Snow, 1978; Sternberg & Weil, 1980). If examinees' performance differ within an achievement test, then it is unlikely that the test will provide a uniform measure of achievement.

Third, psychometric models do not explain test performance. Rather, test performance is explained and examined through an external validation process where the test score interpretation, not the accuracy of the psychometric model, is evaluated. To underscore this limitation, Snow and Lohman (1989) quote Ebel (1962), who noticed this problem 30 years ago: "If more [a priori] systematic and standardized processes of test production could be developed and used [rather that statistical transformations and elaborations of the post hoc test score data], our educational measurements should become not only more consistently reproducible, but what is perhaps even more important, they should become more meaningful" (p. 22).

Achievement tests are cognitive problem-solving tasks. However, the psychometric models used to create these tests have a weak cognitive foundation because they fail to identify the cognitive processes needed to solve test items, they make simplistic assumptions about the psychology of test performance, and they rely on a process external to the model to validate the meaning of the test.

Fortunately, researchers are beginning to address the psychological limitations in psychometric models by combining cognitive and psychometric theory to create new models (see Embretson, 1985; Gustafsson, 1984; Carroll, 1985). It appears that cognitive psychology has influenced and will help improve the models in educational measurement. Equally important is the need for cognitive psychology to influence and improve the procedures used to create tests, as the majority of test developers still rely on Bloom's taxonomy to classify students' cognitive processes.

A Summary and Critique of the Cognitive Domain in the Taxonomy of Educational Objectives

The learning outcomes measured by an achievement test should reflect the objectives of instruction. One way of defining the cognitive objectives of instruction is to use the cognitive domain from the Taxonomy of Educational Objectives. Although many other cognitive classification systems are available (e.g., Gagne, 1977, 1984; Stahl & Murphy, 1981; Miller, Williams, & Haladyna, 1978; Guilford, 1967; Melton, 1964; Tiemann & Markle, 1973; Biggs & Collis, 1982), Bloom's taxonomy remains one of the most popular because it focuses on a comprehensive list of thinking skills commonly used in education and it provides a standard vocabulary for classifying learning outcomes. Bloom's taxonomy was also one of the first cognitive classification systems popularized in education (Gronlund, 1982, 1991; Osterlind, 1989; Smith, 1984; Roid & Haladyna, 1982).

A brief summary of the purpose, organizational principles, and development for Bloom's taxonomy is presented followed by a review of the literature highlighting its strengths and weaknesses.

The purpose of the taxonomy is to classify behaviors that represent the intended outcomes of education. Intended outcomes include all observable behaviors that result from instruction.

These behaviors are assumed to occur in all subject areas and across all grade levels. Tests are used to measure how well the intended outcomes have been learned. Bloom et al. (1956) hoped that the taxonomy would be used by a wide range of educators such as teachers, curriculum planners, and test developers, to to discuss curriculum and evaluation issues with greater precision.

The idea for developing a classification model originated at the 1948 American Psychological Association Conference in Boston. A group of psychometricans agreed to work together and develop a classification framework so they could exchange testing ideas and promote research on the relation between testing and education. A committee was formed (the authors of the taxonomy under the direction of editor Benjamin S. Bloom), and the cognitive domain was completed by 1956.

During the planning stage, it was decided that the taxonomy would be an educational-logical-psychological classification system. The priority in creating the taxonomy was that it be relevant in education and that it be used by the educational community. A logical system meant that terms in the taxonomy would be consistently used and concise. Ambiguous concepts such as knowledge and comprehension were operationally defined and given more precise meaning. The psychological emphasis was included so that the taxonomy would be consistent with the relevant and accepted psychological theory of the time. Also, the developers stressed that the taxonomy would be value-free so that behaviors in any educational institution, and with any educational philosophy, could be objectively classified.

Once the organizational principles were agreed upon, the committee began the process of developing the taxonomy. A large list of educational objectives was collected, the intended behaviors were identified, and groups of similar behaviors were created. From these groups, six main levels of cognitive processes emerged. Each level had several categories. The six major levels, arranged from the simplest processing skill to the most complex, were: (a) knowledge; (b) comprehension; (c) application; (d) analysis; (e) synthesis; and (f) evaluation. This arrangement was believed to be a cumulative hierarchical order where each level was built upon the cognitive skills found in the previous level. Unfortunately, very little information about the criteria used to select, classify, and order the educational objectives is provided by Bloom et al. (1956). The definition for each level is provided in Table 1.

Bloom's taxonomy has three noteworthy strengths. First, it helped educators recognize that knowledge (i.e, recalling information) should not be the only cognitive skill emphasized during education. Bloom et al. (1956) stressed that a range of cognitive skills, as outlined in the taxonomy, should be

Table 1

The Six Major Levels in the Cognitive Domain of the Taxonomy of Educational Objectives

1. Knowledge. Knowledge is defined as remembering previously learned material. This may involve the recall of a wide range of material, from specific facts to complete theories, but all that is required is to bring to mind the appropriate information.

2. Comprehension. Comprehension is defined as the ability to grasp the meaning of material. This may be shown by translating material from one form to another, by interpreting material, and by estimating trends.

3. Application. Application refers to the ability to use learned material in a new and concrete situation. This may include the application of such things as rules, methods, concepts, principles, laws, and theories.

4. Analysis. Analysis refers to the ability to break down material into its component parts so that its structure may be understood. This may include the identification of the parts, analysis of the relations between parts, and recognizing the organizational principles involved.

(continued)

5. Synthesis. Synthesis refers to the ability to put parts together to form a new whole. This may involve the production of unique communication, a plan of operations, or a set of abstract relations.
6. Evaluation. Evaluation is concerned with the ability to judge the value of material for a given purpose. The judgments are based on definite criteria that may be external, internal, created, or provided to the students.

Note. From N. E. Gronlund (1991) summary of the Taxonomy of Educational Objectives: Cognitive Domain.

developed through schooling. Second, the taxonomy contained six unique cognitive processes where each process was operationally defined and differentiated from the others. The taxonomy included examples to demonstrate how each process could be measured in a wide range of subject areas. Third, the taxonomy was widely used. It has been used for describing the educational objectives of courses, curricula, and tests; planning courses and instruction; developing tests at the classroom, school, district, and provincial level; and conducting research on educational outcomes (Furst, 1981). De Landsheere (1977) concluded: "The enormous influence exercised by their [Bloom et al.] imperfect tool proves that it answered a deep and urgently felt need" (p. 105).

Despite these positive contributions, Bloom's taxonomy has also been criticized. Four weaknesses have often been cited. First, the taxonomy has been criticized for being insensitive to the wide range of cognitive processes that can be elicited from a test item. Ormell (1974) found that math problems in senior high school were generally classified as synthesis, but within this cognitive level students performed a diverse range of strategies and operations to formulate their solutions. Consequently, the cognitive processes embedded within each level often cannot be accurately classified with the taxonomy.

Second, many of the assumptions in the taxonomy lack empirical support. For example, Seddon (1978) reviewed several studies and concluded that there is little support for the notion of a cumulative hierarchy in the taxonomy. Only knowledge, comprehension, application, and analysis appear to be hierarchially related to one another (Kropp & Stoker, 1966; Madaus, Woods, & Nuttall, 1973). However, the robustness of this relation across different subject areas has not been adequately studied (Furst, 1981). The lack of empirical support for a hierarchical structure has also been attributed to other weaknesses in the taxonomy such as overlap between the levels, difficulty in reliably classifying student responses, and the possibility that the structure of the taxonomy may be content-dependent (i.e., mathematics and social studies may have a different ordering of levels).

Third, the notion that cognitive processes function independently from both content and affect has been viewed as simplistic and naive (Furst, 1981; Ormell, 1974; Hirst, 1974). These critics argue that much is lost if a only a cognitive approach is used to define the objectives of education. By omitting cognitive-affective processes from the classification system, key skills such as responsibility, group problem-solving, and moral reasoning cannot be measured (Furst, 1981).

Fourth, the idea of creating a value-free classification system has outraged many philosophers and researchers (Ormell, 1974; Hirst, 1974; Seddon 1978; Pring; 1971; Sockett, 1971). The opponents of this idea stress that by emphasizing certain educational outcomes, such as cognitive skills, and diminishing the significance of others, such as attitudes or moral development, values have been imposed. As a result, many consider the taxonomy to be value-laden.

A brief summary of the literature on Bloom's taxonomy highlights the purpose, organizational principles, and development of this classification model. Bloom et al. (1956) have been praised for popularizing the idea that a diverse range of cognitive processes should be included in the objectives of education and for operationally defining six cognitive processes that could be assessed. The taxonomy has been widely used. Bloom's taxonomy has also been criticized. It cannot be used to measure all of the cognitive processes embedded within each level; it has been inadequately studied as there is little empirical support for several key assumptions, such as the cumulative hierarchical structure of the model; it has be criticized for omitting educational objectives such as attitudes; and it is said to be value-laden. Despite these criticism, Bloom's taxonomy remains one of the most popular models for classifying cognitive processes and is frequently used in the table of specifications of achievement tests.

Preview of the Present Study

The time has come to evaluate the procedures used in test development. The purpose of this study was to determine whether Bloom's taxonomy provides an accurate model to guide item writers for anticipating the cognitive processes used by elementary school students to solve items on a large-scale achievement test in mathematics. The cognitive processes are outlined in the table of specifications. The validity of the table depends on two assumptions that were tested in this study: (a) that students use the cognitive processes anticipated by item writers in the same proportions as outlined in the table of specifications (i.e., 3 knowledge items answered by 30 students will elicit 90 knowledge responses), and (b) that high and low math achievers use the cognitive processes anticipated by item writers in the same proportions as outlined in the table of specifications. To investigate the first assumption examinees were asked to think aloud as they solved items on a shortened math achievement test. All responses were tape-recorded. A protocol coding system, based on Bloom's taxonomy, was used to classify examinees' cognitive processes on the basis of their verbal reports. To investigate the second assumption, examinees were classified as either high or low math achievers and compared to evaluate whether the two groups used the cognitive processes anticipated by item writers in the same proportions as outlined in the table of specifications.

Method

Subjects

The sample consisted of 30 Grade 7 students (16 males, 14 females) in the Edmonton Catholic School district. Median age (in yr:mo) was 12:7 (range 12:1 to 14:0). Students were chosen from three classrooms where the teachers agreed to help with the study. Of the 82 parental permission forms distributed in the three classes, 43 were returned (52.4%). Thirty-seven students agreed to participate and 6 declined. Seven students who agreed to participate were not tested because of absenteeism, prior commitments (e.g., class exams), or school events (e.g., dance, special event day).

The performance of students starting Grade 7 was used to approximate the performance of students finishing Grade 6 (the 1991 Achievement Test in Mathematics was administered by Alberta Education in June). The use of Grade 7 students was expected to yield comparable results to students completing Grade 6 as teachers generally believe that summer holidays result in a decrease in academic achievement (Horst, 1976). The Grade 7 students were tested in November. By this time, teachers would have reviewed math concepts and materials from the previous year thereby overcoming the decrease in achievement associated with summer forgetting.

Materials

Mathematics Achievement Subtest (MAS). The Mathematics Achievement Subtest (MAS) is an 18-item multiple-choice exam. The dimensions in the table of specifications and the test items for the MAS were selected from previous achievement tests in mathematics used by Alberta Education. Three cognitive levels across two content areas were measured with the MAS. The cognitive processes assessed were knowledge, comprehension, and application, and were the same cognitive skills measured with the 1991 Achievement Test in Mathematics. The definitions used by Alberta Education to describe these cognitive skills closely resemble the first three levels in Bloom's taxonomy. As a result, Bloom's taxonomy was used to code student's think-aloud protocols.

Knowledge was defined by Alberta Education (1992b) as recognizing or recalling math facts, definitions, rules, procedures, and performing routine math manipulations. Knowledge was defined by Bloom et al. (1956) as the "remembering, either by recognition or recall, of ideas, material, or phenomena" (p. 62). Because both definitions focus on recall and recognition, knowledge was operationally defined as recalling the math operation or solution without calculating an answer. An example of a knowledge item is: In the number 2547.396, the digit 3 is in the (a) ones place, (b) tens place, (c) tenths place, or (d) hundredths place (Alberta Education, 1992b). In this question, students are expected to locate the digit three and recall that the first position to the right of the decimal point is the tenths place.

Comprehension was defined by Alberta Education (1992b) as understanding mathematical principles and concepts and being able to demonstrate this understanding. It may include translating information into different representations such as from numbers to words. Comprehension was defined by Bloom et al. (1956) as an "understanding of the literal message contained in a communication" (p. 89) that could be demonstrated by manipulating, interpreting, and explaining concepts and ideas. Both definitions require students to demonstrate their understanding of a concept. As as result, comprehension was operationally defined as performing the math operation required in the question and generating a solution. An example of a comprehension item is: Find the product. 3.23 x 0.9 = ; (a) 2.787, (b) 2.907, (c) 27.87, or (d) 29.07. In this question, students are expected to multiply the decimal numbers together and put the decimal in the correct place in the answer (Alberta Education, 1990).

Application was defined by Alberta Education (1992b) as solving math problems by using previously learned skills and knowledge. Application was defined by Bloom et al. (1956) as the ability to "apply the appropriate abstraction without having to be prompted as to which abstraction is correct or without having to be shown how to use it in that situation" (p. 120). In other words, application is the use of previously learned materials in new situations. Both definitions emphasize using previously learned knowledge to solve problems. Application was operationally defined as performing the math operations required in the question, generating an intermediate solution, and then applying the intermediate solution to reach the final answer. An example of an application item is: Carol correctly answered 6 out of 8 questions. What was her score in percent?; (a) 48%, (b) 68%, (c) 70%, or (d) 75%. In this question, students are required to convert 6/8 to a percentage (Alberta Education, 1990) by dividing 8 into 6 to get an intermediate solution of 0.75, and then multiplying the intermediate solution by 100 to produce the final solution.

The content areas on the MAS were numeration and operations and properties. These two areas represent 52.7% of the content coverage on the 1991 Achievement Test in Mathematics. Numeration includes concepts such as recognizing and manipulating mathematical patterns, place values, numbers (whole, decimal, fractions), and numerical relationships (comparing, ordering, rounding). Operations and properties includes concepts such as applying number properties (commutative, associative, distributive) as well as adding, subtracting, multiplying, and dividing whole numbers and decimals (Alberta Education, 1992a).

Combining the three cognitive levels with the two content areas produced the table of specifications used for the MAS in this study (see Figure 2). The table contained an equal number of items in each cell with a total of 6 items for knowledge, comprehension, and application and 9 items for both numeration and operations and properties. To ensure that the MAS would not be confounded by either the math concepts within each content area or the item difficulty (the concepts and difficulties for each item are provided in 1991 Grade 6 Mathematics Achievement Test Information Sheet), items in each cell were selected so that they measured different math concepts and had a range of item difficulties (varying from .34 to .65). Because these conditions could not be fulfilled with items from the 1991 exam, two items from both the 1987 and the 1983 Mathematics Achievement Test were used. The earlier math achievement tests contained the same table of specifications as 1991 exam, and all three achievement tests were based on the 1982 math curriculum (Alberta Education, 1992b).

Items for the MAS were ordered unsystematically with the constraints that the three cognitive levels and the two content areas were divided evenly among the first and second half of the test. The order of the items in the first form was reversed to to create a second form. The MAS is provided in the Appendix.

Figure 2. The Table of Specifications for the Mathematics Achievement Subscale (MAS). Each test item in a cell was formed by combining one cognitive level and one content areas. Each cell contains three items. The MAS item number, the item difficulty calculated by Alberta Education, and the year the item was used is provided.

Cognitive Level

Content Area	Knowledge	Comprehension	Application
Numeration	3 (.377) 1991	1 (.626) 1991	9 (.480) 1991
	5 (.649) 1991	12 (.344) 1991	10 (.633) 1983
	7 (.443) 1991	16 (.497) 1991	14 (.554) 1991
Operations & Properties	11 (.590) 1991	4 (.384) 1991	2 (.600) 1991
	13 (.596) 1983	8 (.577) 1991	6 (.530) 1991
	15 (.650) 1987	18 (.430) 1991	17 (.482) 1987

Procedure

Students within each of the three classes were rank-ordered into two achievement categories according to math grades obtained for the first reporting period. Of the 30 students who agreed to participate, 15 students (9 males, 6 females) from the top half, and 15 students (7 males, 8 females) from the bottom half of each class were tested. Across the three classes, the mean math grades for the low scorers in the high achievement group, computed from teacher-assigned unit math exams, were 9% to 14% greater than the mean math grades for high scorers in the low achievement group. To minimize possible effects due to item order, test order was counterbalanced as evenly as possible for both males and females.

Participants were individually tested in an empty classroom. Students were asked to think aloud as they solved each test item and to say all the thoughts and strategies that came to mind as they formulated their solution. After one of four possible multiple-choice options was selected, students were asked to explain why they chose that option. All responses were tape-recorded. Three practice items were completed prior to beginning the MAS.

Results

Mathematics Achievement Subscale

The first set of analyses was performed to compare students' performance on items from Mathematics Achievement Subscale (MAS) with items from the three provincial achievement exams, and to assess the reliability and validity of the MAS. The total score, expressed in percentage, on the MAS was comparable to the score on the 18 items taken from the three provincial achievement tests in mathematics (65.0% vs. 61.0%, respectively), indicating that students in the sample performed in a similar manner to students in the province, and that the 18 test items produced similar scores. Cronbach's alpha for the 18-item MAS was .81, indicating a reasonably high degree of internal consistency. The MAS and teacher-assigned grades were strongly correlated, $\underline{r}(30) = .88$, $\underline{p} < .01$, supporting the validity of the MAS as a measure of math achievement.

Inter-rater Agreement

To ensure that student's cognitive processes were consistently coded, inter-rater agreement was assessed. One rater, who had no experience with the study, was trained to classify the cognitive processes in student's responses. The scoring system was described to the rater; responses from two randomly-selected students were coded for practice; and responses from five randomly-selected students (16.7% of the sample) were then coded by the rater and compared to the coding of the researcher to assess inter-rater agreement. Of the 90 responses coded, 75 agreements occurred (83.3%), indicating that student's cognitive processes were consistently coded.

Disagreements were evenly distributed across items in the three cognitive levels: Five disagreements occurred with knowledge items, 6 with comprehension items, and 4 with application items.

The Cognitive Processes Used by Students Overall

To investigate the first research question--do students use the cognitive processes identified by item writers in the same proportions as outlined in the table of specifications?--chi-square tests were conducted on the marginal frequencies overall and across the two content areas of the MAS contingency table, as presented in Table 2. Chi-square tests were used to highlight differences between the responses expected by Alberta Education and observed from students. Only the frequencies in the summary matrices (matrices 7 to 9) are discussed in this section. The row and column entries are the expected and the observed cognitive responses, respectively. Guesses (66 responses) and unclassified cognitive processes (6 responses) were excluded from all analyses.

Of the 468 cognitive responses reported by the students, 251 (53.6%) matched the cognitive levels anticipated by the item writers. When the marginal values for matrix 9 were examined, the expected responses for the three cognitive levels (row sums: 158, 159, and 151) differed significantly from the observed values (column sums: 89, 346, and 33), X^2 (2, $\underline{N}$ = 30) = 342.28, $\underline{p} < .01$, indicating that students used knowledge, comprehension, and

Table 2

The Mathematics Achievement Subtest Contingency Table Containing the Frequencies for the Expected (Rows) and Observed (Columns) Cognitive Responses as a Function of Achievement Group and Content Area

Content Area		Achievement: High Know	High Comp	High Appl	Low Know	Low Comp	Low Appl	Observed Marginal Sums		
Numeration	Know	10	32	1	12	30	0	22	62	1
	Comp	0	43	0	0	39	0	0	82	0
	Appl	[1] 0	25	16	[3] 0	33	6	[7] 0	58	22
Operations & Properties	Know	28	12	0	19	14	0	47	26	0
	Comp	5	35	0	4	33	0	9	68	0
	Appl	[2] 6	28	7	[4] 5	22	3	[8] 11	50	10
Expected Marginal Sums		38	44	1	31	44	0	69	88	1
		5	78	0	4	72	0	9	150	0
		[5] 6	53	23	[6] 5	55	9	[9] 11	108	32

Note. The number used to reference each matrix in the manuscript appear in the bottom left corner of each cell.

application processes to solve items on the MAS, but not in the same proportions as identified by item writers. Across the two content areas, a similar result occurred as the cognitive responses expected by item writers failed to match the cognitive responses observed from students in numeration (the marginal values in matrix 7 are 85, 82, 80 for the rows vs. 22, 202, 23 for the columns), X^2 (2, N = 30) = 262.92, p < .01, with only 126 out of 247 (51.0%) response matches, and in operations and properties (the marginal values in matrix 8 are 73, 77, 71 for the rows vs. 67, 144, 10 for the columns), X^2 (2, N = 30) = 111.20, p < .01, with only 125 out of 221 (56.6%) response matches. These results reinforce the conclusion that students, in general, used the processes described in Bloom's taxonomy to solve MAS items, but not in the same proportions as identified by item writers when summarized overall (i.e., matrix 9) or when summarized as a function of the two content area (i.e., matrices 7 and 8).

The Cognitive Processes Used by High and Low Achievers

To investigate the second research question--do high and low math achievers use the cognitive processes anticipated by item writers in the same proportions as outlined in the table of specifications?--chi-square, analysis of variance, and loglinear tests were used to evaluate the data.

First, to demonstrate that the sample was composed of two distinct achievement groups, mean scores on the MAS were calculated. The overall mean scores differed for the high and low mathematics achievers [mean = 14.9, s.d. = 1.6 vs. mean = 8.5, s.d. = 2.7, for high and low achievers, respectively, $t(28) = 8.01$, $p < .01$], demonstrating that the performance of two achievement groups was different on the MAS.

Earlier, chi-square tests were used to compare the marginal frequencies of matrices 7, 8, and 9 in Table 2. It was noted that the frequencies between expected and observed responses differed overall, and in each content area. When cognitive level was partitioned into high and low achievers, the chi-square test statistic remained significant for both high achievers [the marginal values in matrix 5 are 83, 83, 82 for the rows vs. 49, 175, 24 for the columns), X^2 (2, N = 15) = 156.93, $p < .01$, as only 139 out of 248 (56.0%) responses matched], and low achievers [the marginal values in matrix 6 are 75, 76, 69 for the rows vs. 40, 171, 9 for the columns), X^2 (2, N = 15) = 187.26, $p < .01$, as only 112 out of 220 (50.9%) responses matched]. These results suggest that students of different abilities use similar cognitive processes to solve MAS questions, but not in the same proportions as intended by the item writers.

Changes in agreement between the expected and the observed cognitive responses were analyzed with a 2 (Achievement: High vs. Low) x 2 (Sex) x 3 (Cognitive Level: Knowledge vs. Comprehension vs. Application) x 2 (Content Area: Numeration vs. Operations and

Properties) analysis of variance with repeated measures on the last two variables. Each student's response was scored by comparing it to the cognitive process anticipated by the item writers. Agreements were coded 1 and disagreements 0. Scores could ranged from 0 (no match) to 3 (perfect match), as each cognitive level/content area cell in the table of specifications for the MAS contained three items (see Figure 2). One student was omitted from this analysis because he did not complete an item on the MAS. Mean scores are presented in Table 3.

The mean match scores differed for the two achievement groups, $F(1, 25) = 9.40$, $p < .01$, $MSe = .38$, and for the three cognitive levels, $F(2, 50) = 143.59$, $p < .01$, $MSe = .41$. These main effects were qualified by a Cognitive Level x Content Area, $F(2, 50) = 24.60$, $p < .01$, $MSe = .28$, as well as an Achievement x Cognitive Level x Content Area interaction, $F(2, 50) = 5.28$, $p < .01$.

For knowledge, the match between the expected and the observed cognitive responses in numeration did not differ for the high and low achievers, but in operations and properties it did, as the high group had more matches, $F(1, 50) = 8.09$, $p < .01$. Both high [$F(1, 100) = 34.62$, $p < .01$] and low achievers [$F(1, 100) = 4.45$, $p < .05$] answered more operations and properties items using the expected cognitive processes than numeration items.

Table 3

Mean Agreement Scores For Expected and Observed Responses on Knowledge, Comprehension, and Application as a Function of Achievement Group and Sex

	Achievement Group		Sex	
Cognitive Level/Content Area	High	Low	Male	Female
Knowledge/Numeration	.67	.86	.73	.79
Knowledge/Operations & Properties	1.87	1.29	1.47	1.71
Comprehension/Numeration	2.87	2.57	2.80	2.64
Comprehension/Operations & Properties	2.33	2.29	2.20	2.43
Application/Numeration	1.07	.43	.80	.71
Application/Operations & Properties	.47	.21	.33	.36

Note. A higher mean score corresponds to more agreement between the expected and the observed cognitive response.

For comprehension, mean ratings in both content areas did not differ between the high and low achievers. However, high achievers answered more items as anticipated by Alberta Education in numeration than in operations and properties, $\underline{F}(1, 100) = 7.01$, $\underline{p} < .01$, whereas the agreements across the content areas did not differ for the low achievers.

For application, mean match in numeration differed between the high and low achievers, $\underline{F}(1, 50) = 9.85$, $\underline{p} < .01$, but not in operations and properties. High achievers answered more items as expected in numeration than in operations and properties, $\underline{F}(1, 100) = 8.65$, $\underline{p} < .01$, but the responses across content area did not differ for the low achievers. The ANOVA summary table is presented in Table 4.

Three general findings emerge from these analyses. First, Alberta Education designed the cognitive levels in the table of specifications to be mutually exclusive (i.e., one item per cognitive level/content area). However, when mean match score was used as the dependent variable, a three-way interaction was found (a two-way interaction of cognitive level and content area also occurred, but only the highest-order interaction is interpreted here). This result indicates that items were solved with a variety of the cognitive processes listed in the table of specifications and that achievement group, cognitive level, and content area influence item classification. Second, comprehension

Table 4

Summary Table for the 2 (Achievement) x 2 (Sex) x 3 (Cognitive Level) x 2 (Content Area) Analysis of Variance With Repeated Measures on the Last Two Variables

Source	df	MS	F
Between Subject Variables			
Achievement	1	3.56	9.40*
Sex	1	.41	1.08
Achievement x Sex	1	.47	1.24
Error	25	.38	
Within Subject Variables			
Cognitive Level			
Cognitive Level	2	58.97	143.59*
Sex x Cognitive Level	2	.08	.20
Achievement x Cognitive Level	2	.28	.68
Sex x Achievement x Cognitive Level	2	.60	1.46
Error	50	.41	

(continued)

Source	df	MS	F
Content Area			
Content Area	1	.02	.06
Sex x Content Area	1	.67	1.84
Achievement x Content Area	1	.07	.19
Sex x Ability x Content Area	1	.88	2.42
Error	25	.37	
Cognitive Level x Content Area			
Cognitive Level x Content Area	2	6.86	24.60*
Sex x Cognitive Level x Content Area	2	.10	.37
Achievement x Cognitive Level x Content Area	2	1.47	5.28*
Sex x Achievement x Cognitive Level x Content Area	2	.19	.69
Error	50	.28	

*$p < .01$

had the highest mean match score suggesting that it is the cognitive process most easily anticipated by item writers. Bloom et al. (1956) foreshadowed this finding when they speculated that comprehension processes were "the largest general class of intellectual abilities and skills emphasized in school" (p. 89). It appears that students use comprehension processes to solve math problems. Conversely, knowledge and application processes were poorly anticipated by the item writers. Third, the two achievement groups tend to use similar cognitive processes to solve test items. However when differences occurred, the high achievers had more matches between the expected and observed responses than low achievers. This finding suggests that item writers are more accurate at anticipating the cognitive processes used by high math achievers than by low math achievers.

To examine the achievement group, content area, and cognitive level concurrently, a loglinear analysis was conducted on the cell frequencies in matrices 1 to 4. The goal in creating the loglinear model was to select a combination of parameters that produced a nonsignificant, interpretable X^2 coefficient. In this context, a nonsignificant result indicates that the chosen parameters yield frequencies that provide an acceptable fit to the contingency table data. Although many different models were tested, only the most parsimonious model (i.e., fewest parameters and interpretable) is reported (Feinberg, 1980). The chosen model contained an achievement group main effect and a content area x cognitive level interaction, X^2 (17, N = 30) = 9.91, p = .91. The results of the loglinear analysis were used to investigate the misclassified cognitive responses in matrix 9, as reported in the next section.

The diagonal elements of matrix 9 in Table 2 contain matches between the cognitive responses expected by Alberta Education and observed from students, and the off-diagonal elements contain mismatches. Matrix 9 contains two large off-diagonal elements that represent discrepancies between the responses expected by Alberta Education and the responses observed from students: 88 responses were expected to be knowledge but observed as comprehension and 108 responses were expected to be application but observed as comprehension. To understand the misclassified responses, chi-square tests were used to evaluate the cell frequencies from the content area x cognitive level interaction, as suggested by the results of the loglinear analysis. For the 88 responses Alberta Education expected to be classified as knowledge but were observed as comprehension, there was a difference between the content areas [the expected cell frequencies in summary matrices 7 and 8 are 44, 44 (obtained from 88/2) vs. the observed cell frequencies of 62, 26, X^2 (1, N = 30) = 14.73, p < .01], as the cognitive processes required to solve numeration items were misclassified more frequently than for operations and properties items. For the 108 responses Alberta Education expected to be classified as application but were observed as comprehension, there was no difference between the two content areas [the expected cell frequencies in summary matrices 7 and 8 are 54, 54 (obtained from 108/2) vs. the observed cell frequencies of 58, 50), X^2 (1, N = 30) = .59, p = .44]. Differences between the high and low achievers were not statistically significant, as both groups contributed equally to the 88 and the 108 misclassified responses.

In summary, the parameters achievement group and content area x cognitive level were required to fit a loglinear model to the contingency table data. Based on this finding, post-hoc analyses were used to evaluate the cell frequencies in the content area x cognitive level interaction. For the 88 responses Alberta Education expected to be classified as knowledge but were observed as comprehension, there was a difference between the content areas as the cognitive processes required to solve numeration items were misclassified more often that for operations and properties items. Conversely, for the 108 responses Alberta Education expected to be classified as application but were observed as comprehension, there was no significant difference between the misclassified cognitive processes for numeration and operations and properties items. High and low achievers contributed equally to both the 88 and 108 misclassified cognitive responses.

Evaluating the Knowledge and Application Items Coded as Comprehension

Student protocols were evaluated to address the question, why are many items in matrix 9 of Table 2 that were expected to be knowledge and application observed as comprehension?

Items expected to be knowledge, but observed as comprehension. Of the six knowledge items in the MAS, items 5, 7, and 15 (see Appendix) contain 70 of the 88 misclassified responses (79.5%), as shown in Table 5. Numeration items (in particular, items 5 and 7) were the most frequently misclassified, as would be expected from the results of the loglinear analysis. When the response frequencies of high and low achievers were examined, the same three knowledge items accounted for a majority of the misclassified responses for both groups. Items 5, 7, and 15 accounted for 34 of the 44 (77.3%) and 36 of the 44 (81.8%) misclassified responses for high and low achievers, respectively. Because the same three items were frequently misclassified across the knowledge category and the achievement groups, students' protocols were examined on these items.

Both Alberta Education (1992b) and Bloom et al. (1956) differentiate knowledge and comprehension. Knowledge was described as remembering, either by recognition or recall, mathematical ideas and material. For this study, knowledge was operationally defined as recalling the math operation or solution

Table 5

Summary of Student Protocol Responses as a Function of the Cognitive Process Expected by Alberta Education and Observed from High and Low Math Achievers

Test Item	Expected	Observed: High[a]				Observed: Low[b]			
Item #3	K (N)[c]	K-10[d]	C-5	A-0	G-0	K-12	C-1	A-0	G-1
Item #5	K (N)	K-0	C-15	A-0	G-0	K-0	C-15	A-0	G-0
Item #7	K (N)	K-0	C-12	A-1	G-2	K-0	C-14	A-0	G-1
Item #11	K (O/P)[e]	K-12	C-3	A-0	G-0	K-10	C-3	A-0	G-2
Item #13	K (O/P)	K-13	C-2	A-0	G-0	K-8	C-4	A-0	G-3
Item #15	K (O/P)	K-3	C-7	A-0	G-4	K-1	C-7	A-0	G-7
Item #1	C (N)	K-0	C-15	A-0	G-0	K-0	C-14	A-0	G-1
Item #12	C (N)	K-0	C-15	A-0	G-0	K-0	C-13	A-0	G-2
Item #16	C (N)	K-0	C-13	A-0	G-1	K-0	C-12	A-0	G-2
Item #4	C (O/P)	K-0	C-13	A-0	G-2	K-1	C-14	A-0	G-0
Item #8	C (O/P)	K-5	C-10	A-0	G-0	K-3	C-12	A-0	G-0
Item #18	C (O/P)	K-0	C-12	A-0	G-3	K-0	C-7	A-0	G-8
Item #9	A (N)	K-0	C-15	A-0	G-0	K-0	C-14	A-0	G-1
Item #10	A (N)	K-0	C-6	A-7	G-2	K-0	C-9	A-3	G-3
Item #14	A (N)	K-0	C-4	A-9	G-2	K-0	C-10	A-3	G-2
Item #2	A (O/P)	K-0	C-14	A-0	G-0	K-0	C-8	A-0	G-7
Item #6	A (O/P)	K-6	C-6	A-0	G-3	K-5	C-3	A-1	G-5
Item #17	A (O/P)	K-0	C-8	A-7	G-0	K-0	C-11	A-2	G-2

Note. K-Knowledge; C-Comprehension; A-Application; and G-Guess. Unclassified responses are not included.

[a]*n* = 15. [b]*n* = 15. [c]Numeration item. [d]Response mode is underlined for each test item. [e]Operations and Properties item.

without calculating the answer. Comprehension was described as understanding mathematical principles and concepts and being able to demonstrate this understanding. Comprehension was operationally defined as performing the math operation required in the question and generating a solution.

For item 5, there was little response variability as 29 students solved the problem directly. The numbers in parentheses were multiplied together, and the products were added to generate the solution. The salient cognitive processes were multiplication and addition, and the calculations were often performed in the test booklet. These processes seem to demonstrate an understanding of standard notation, and go beyond recognizing or recalling a solution.

Item 7 elicited a variety of strategies. Nine students first counted the total number of pieces in the pie, then counted the number of shaded pieces and formed a fraction of shaded to unshaded pieces. Students then reduced the fraction using the lowest denominator and selected the appropriate solution. The calculations were often performed in the test booklet. A second strategy, used by 5 students, was to create the fraction by counting, but to explain that if two pieces of pie were counted as one the correct solution would be found. No student recalled the solution.

Item 15 produced the most response variability as seven different strategies were identified. The most common approach, used by three students, was to identify the counting ratio, locate some marker points, and count. One examinee explained: "You can count in 2s. 50% is 10, 100% is 20, so 90% is 18." Two students recalled that 90% was .9, and then computed in the test booklet that 20 x .9 = 18. Another strategy, used by two students, was to set up the problem as 90/100 = x/20, and identify a common divisor through trial-and-error. Five was often used and applied to both the numerator and denominator so that 100/5 = 20, and 90/5 = 18. Knowledge was another strategy used to solve item 15, as three students explained: "18 out of 20 is 90% because I've got 90% on tests before." The knowledge response clearly demonstrates recalling the solution. In contrast, the comprehension responses involve calculating the solution.

By examining three misclassified items, two findings emerged. First, there was response variability as students used a variety of different strategies to solve an item. Second, many knowledge items were solved with processes and strategies that corresponded to comprehension, as defined in this study.

Items expected to be application, but observed as comprehension. Of the six application items in the MAS, items 2, 9, and 17 (see Appendix) contain 70 of the 108 misclassified responses (64.8%), as shown in Table 5. Numeration item 9 and operations and properties item 17 were the most frequently misclassified. When the response frequencies of high and low achievers were examined, the same three application items accounted for a majority of the misclassified responses. Items 2, 9, and 17 accounted for 37 of the 53 (69.8%), and 33 of the 55 (60.0%) misclassified responses for high and low achievers, respectively. Because the same three items were frequently misclassified, students' protocols were examined on these items.

Both Alberta Education (1992b) and Bloom et al. (1956) differentiate comprehension and application. Comprehension is described in the previous section. Application is described as using previously learned materials to solve problems in new situations. Application was operationally defined as performing the math operations required in the question, generating an intermediate solution, and then applying the intermediate solution to reach the final answer. For item 2, 19 students directly calculated their answer by either dividing or multiplying and their response was coded as comprehension. For example, 16 students moved the decimal one place in both numbers and then divided 245 by 35. Three students multiplied 24.50 by 3.5 to get the incorrect answer of 85.75.

For item 9, 28 students used the information in the question to calculate the solution, using either paper-and-pencil or mental computations. As a result, the majority of student responses were coded as comprehension. A typical student protocol was: "The hundreds place is 100; the tens and ones are 0; the tenths is (pause) 0.2, as 2 x 1 = 2; and the hundredths is (pause) .08, as 2 x 4 = 8. The number is 100.28." As in item 2, all of the information needed to calculate the solution for item 9 was in the question. Therefore, students were able to calculate the answer directly.

For item 17, the majority of students who answered the item with a comprehension strategy used division. However, the unique feature of this item was that comprehension strategies generally led to the wrong answer. Seventeen students divided 80 by 7 to get 11 remainder 3. Based on this calculation, students chose 11 as the final answer, and one student noted: "The answer is 11. Hmmmmm..., 3 people died". Nevertheless, the salient strategy was division as 19 students calculated an answer directly.

By examining three misclassified items that were expected to elicit application processes by more commonly produced comprehension responses, two findings emerged. First, there was response variability as each item elicited several different strategies. Second, application items were readily solved with strategies that could be interpreted as comprehension. Most students used information within the question to directly calculate the answer. However, comprehension strategies consistently produced the incorrect solution for item 17.

Summary

The Mathematics Achievement Subscale (MAS) contained 18 multiple-choice items selected from previous math achievement tests used by Alberta Education. Three cognitive processes (knowledge, comprehension, and applications) in two content areas (numeration and operations and properties) were assessed. Students' total score on the MAS was comparable to the 18 item score from the three provincial achievement tests indicating that students in the sample performed in a similar manner to students in the province, and that the items produced similar scores overall. The MAS was reliable, and scores on the MAS correlated strongly with teacher-assigned grades in mathematics.

The 30 Grade 6 students who participated in the study were instructed to think aloud as they solved items on the MAS. Bloom's taxonomy was used to code students' responses because it is the model most commonly used to assess cognitive processes in the table of specifications of achievement tests (Gronlund, 1991; Osterlind, 1989), and it matches the cognitive processes measured by Alberta Education on the 1991 Achievement Test in Mathematics. Inter-rater agreement has high, indicating that student's cognitive processes were consistently coded.

To address the first research question--do students use the cognitive processes identified by items writers in the same proportions as outlined in the table of specifications?--

response frequencies were calculated and chi-square tests were conducted on the marginal frequencies overall and across the two content areas of the MAS contingency table. The results of these tests indicate that students use knowledge, comprehension, and application processes to solve MAS items, but not in the same proportions overall, or across the two content areas, as indicated in the table of specifications. More than 40% of the cognitive responses observed from students did not match the responses expected by item writers.

To address the second research question--do students of different abilities use the cognitive processes identified by item writers in the same proportions as outlined the table of specifications?--chi-square, analysis of variance, and loglinear tests were used to evaluate the data. High and low math achievers were compared. The two groups differed on their teacher-assigned grades in mathematics and on their overall MAS scores as high achievers outscored low achievers.

When chi-square tests were conducted on the response frequencies of high and low achievers, students in the two groups used similar processes, but in different proportions as outlined in the table of specifications. Changes in agreement between the expected and the observed cognitive responses were assessed with an analysis of variance. Three findings emerged. First, the achievement group x cognitive level x content area interaction demonstrated that the match between expected and observed responses relies on a complex relationship between these variables. The loglinear analyses confirmed this finding as a content area x cognitive level parameter was needed to adequately fit a model to the data. The interaction between achievement group, cognitive level, and content area indicated that the variables are not mutually exclusive as outlined in the table of specifications. Second, comprehension items had the highest mean match score suggesting that this class of cognitive processes is often used by elementary school students in mathematics. Bloom et al. (1956) alluded to this finding when they postulated that comprehension processes were highly emphasized in school. Compared to comprehension, knowledge and application processes were poorly anticipated by the item writers. Third, although high and low achievers had similar matches between the responses expected by item writers and the responses observed during the think-aloud protocols, the two significant differences favored the high achievers. This finding suggested that item writers are more accurate at anticipating the responses of high achievers.

When the elements in matrix 9 of the MAS contingency table (see Table 2) were examined, two large off-diagonal elements were identified. The off-diagonal elements represent discrepancies between the responses expected by Alberta Education and the responses observed from students. For the 88 responses Alberta

Education expected to be classified as knowledge but were observed as comprehension, there was a difference between the content areas as the cognitive processes required to solve numeration items were misclassified more frequently that for operations and properties items. Conversely, for the 108 responses Alberta Education expected to be classified as application but were observed as comprehension, there was no difference between the misclassified cognitive processes between the numeration and operations and properties items. High and low achievers contributed equally to both the 88 and 108 misclassified cognitive responses. To account for the misclassified cognitive responses, student protocols were examined on several items. Two findings emerged from the protocol analysis. First, there was response variability as students used a variety of different strategies to solve the items. This finding was consistent with Ormell's (1974) critique of Bloom's taxonomy. Ormell showed that math items elicit many different strategies, and that these strategies are lost when coded with the general levels used in Bloom's taxonomy. Second, many knowledge and application items were solved with processes and strategies that could be interpreted as comprehension.

General Discussion

The results of this study indicate that the cognitive domain in the Taxonomy of Educational Objectives (Bloom et al., 1956) does not provide an accurate model to guide item writers for anticipating the cognitive processes used by elementary school students to solve items on a large-scale achievement test in mathematics. The model failed for five important reasons. First, the cognitive processes expected by item writers matched the processes used by students in only 55% of the cases outlined in the table of specifications. This finding demonstrates that Bloom's taxonomy enabled item writers to correctly anticipate the cognitive processes used by students about half of the time. Second, within each level of the taxonomy there was response variability. For example, item 15 on the MAS was solved with seven different strategies. This result demonstrates that the levels in Bloom's taxonomy, as used in test construction, conceal response variability, and that much of the cognitive complexity is lost by coding responses under general headings such as knowledge or application. Furthermore, when think-aloud protocols were used to evaluate student's cognitive processes, many items that were expected to be solve using knowledge and application processes were actually solved using comprehension processes. Third, the table of specifications treats the cognitive levels and content areas as mutually exclusive but when the match between expected and observed responses was examined, it was shown that the achievement groups, cognitive levels, and content areas interact. This finding indicates that students solved items with a variety of cognitive processes in the table of specifications and that achievement group, cognitive level, and content area influence item classification. Fourth, comprehension processes were readily anticipated by item writers, but knowledge and application processes were not. Fifth, item writers were able to infer the cognitive processes used by high math achievers more accurately than by low math achievers. Consequently, the cognition section in the table cf specifications is a more valid guide of the mental processes used by high achieving math students.

If, by using Bloom's taxonomy, item writers are unable to accurately anticipate students' cognitive processes, what can be done to improve this aspect of test design? Three suggestions are provided. First, the cognition section in the table of specifications should *not* be used for test interpretation as it does not accurately identify the cognitive processes used by students to solve test items. By removing the cognition section from test manuals and publications, users would be discouraged from making erroneous interpretations about student's cognitive skills based on their achievement test scores.

Second, there is a need to adopt the concepts and methods of cognitive psychology. For example, rather than trying to select items that are *intended* to measure a cognitive process, item writers should select items that are *known* to elicit a cognitive process. A great deal of research in cognitive psychology is now directed at understanding the cognitive processes used by students to solve math problems (e.g., Bisanz & LeFevre, 1990; Siegler, 1988; Brown & Burton, 1978). To demonstrate how cognitive research could be incorporated into test design, an example is provided.

Brown and Burton (1978) studied student's cognitive processes in subtraction by examining error patterns on problem-solving tasks. They discovered that many of the errors made by students were due to "buggy" computational procedures rather than to carelessness, lack of knowledge about subtraction, or lack of motivation in math. Brown and Burton's research methods could also be used by test developers where "buggy" foils (i.e., the multiple-choice options that are not correct), associated with specific erroneous procedures, could be built into achievement tests. If a student was using an erroneous procedure (e.g., several students in the present study answered item 3 in the following way: 3/4 = 3.4), distinct error patterns would emerge. Once the pattern was identified, remedial instruction would be used to correct the erroneous procedure. Error pattern analysis has two advantages: It focuses on the cognitive processes actually used by the student and it has direct instructional benefits. Item response patterns could also be used to identify students for whom a test is inappropriate (i.e., item bias research) or schools with curricula that do not match the content of an achievement test (Harnisch & Linn, 1981).

Third, protocol data should be collected and analyzed during the pilot phase of test development. When Bloom's taxonomy was developed, the editorial staff collected a large list of educational objectives, identified the intended behaviors, and created groups of similar behaviors resulting in the six cognitive levels. During the development of Alberta Education's 1991 Mathematics Achievement Test, items for each cognitive level in the table of specifications were created by writers who inferred the cognitive processes required by the examinee to answer the item correctly. It both cases, the cognitive processes of students were inferred rather than measured directly. If test developers hope to successfully assess students' cognitive processes, much more research should be conducted with think-aloud protocols to directly evaluate the problem-solving strategies used by students to solve achievement test items.

In summary, only when we eliminate the cognition section from test manuals, adopt the concepts of cognitive psychology, and use the mental processes of students to guide test development will we begin to assess and understand all of the cognitive processes, and their unique organization, necessary for students to solve items in the achievement domain.

References

Alberta Education (1990, September). Student Achievement Testing Program Bulletin: Grade 6 Mathematics. Edmonton, AB: Alberta Education Student Evaluation Branch.

Alberta Education (1992a). Program of studies (elementary schools) ECS - Grade 6. Edmonton, AB: Alberta Education Curriculum Branch.

Alberta Education (1992b). Achievement testing program provincial report: June 1991 administration. Edmonton, AB: Alberta Education Student Evaluation Branch.

Best, J. B. (1989). Cognitive Psychology (2nd ed.). St. Paul, MN: West.

Biggs, J. B., & Collis, K. F. (1982). Evaluating the quality of learning: The SOLO Taxonomy. New York: Academic Press.

Bisanz, J., & LeFevre, J. (1990). Strategic and nonstrategic processing in the development of mathematical cognition. In D. F. Bjorklund (Ed.), Children's strategies: Contemporary views of cognitive development (pp. 213-244). Hillsdale, NJ: Erlbaum.

Bloom, B. S. (Ed.), Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). A taxonomy of educational objectives: Handbook I: Cognitive domain. New York: Longmans, Green.

Brown, J. S., & Burton, R. B. (1978). Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science, 2, 155-192.

Carroll, J. B. (1985). Domains of cognitive ability. Paper presented at the meeting of the American Association for the Advancement of Science, Los Angeles.

De Landsheere, V. (1977). On defining educational objectives. Evaluation in Education, 1, 73-150.

Ebel, R. L. (1962). Content standard test scores. Educational and Psychological Measurement, 22, 15-25.

Ebel, R. L., & Frisbie, D. A. (1986). Essentials of educational measurement (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Embretson, S. E. (1985). Introduction to the problem of test design. In S. E. Embretson (Ed.). Test design: Developments in psychology and psychometrics (pp. 3-17) New York: Academic Press.

Emmerich, W. (1989). Appraising the cognitive features of subject tests (Research Rep. No. RR-89-53). Princeton, NJ: Educational Testing Service.

Fienberg, S. E. (1980). The analysis of cross-classified categorical data (2nd ed.). Cambridge, MA: MIT Press.

Fleishman, E. A., & Quaintance, M. K. (1984). Taxonomies of Human Performance. New York: Academic Press.

Furst, E. J. (1981). Bloom's taxonomy of educational objectives for the cognitive domain: Philosophical and educational issues. Review of Educational Research, 51, 441-453.

Gagne, R. M. (1977). The conditions of learning (3rd ed.). New York: Holt, Rinehart, and Winston.

Gagne, R. M. (1984). Learning outcomes and their effects: Useful categories of human performance. American Psychologist, 39, 377-385.

Gronlund, N. E. (1982). Constructing achievement tests (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Gronlund, N. E. (1991). How to write and use educational objectives (4th ed.). New York, NY: MacMillan.

Guilford, J. P. (1967). The nature of human intelligence. New York: McGraw-Hill.

Gustafsson, J. E. (1984). A unifying model for the structure of intellectual abilities. Intelligence, 8, 179-203.

Harnish, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational Measurement, 18, 133-146.

Hirst, P. H. (1974). Knowledge and the curriculum: A collection of philosophical papers. London: Routledge & Kegan Paul.

Horst, D. P. (1976). What's bad about grade-equivalent scores (ESEA Title I Evaluation and Reporting System Technical Paper No. 1). Mountain View, CA: RMC Research Corporation.

Kropp, R. P., & Stoker, H. W. (1966). The construction and validation of tests of the cognitive processes described in the taxonomy of educational objectives (Cooperative Research Project No. 2117). U.S. Office of Education.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Madaus, G. F., Woods, E. M., & Nuttall, R. L. (1973). A causal model analysis of Bloom's taxonomy. American Educational Research Journal, 10, 253-262.

Melton, A. W. (1964). The taxonomy of human learning: Overview. In A. W. Melton (Ed.), Categories of human learning (pp. 325-339). New York: Academic Press.

Miller, H. G., Williams, R. G., & Haladyna, T. M. (1978). Beyond facts: Objective ways to measure thinking. Englewood Cliffs, NJ: Educational Technology Publications.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 335-366). New York: American Council on Education, Macmillian.

Ormell, C. P. (1974). Bloom's taxonomy and the objectives of education. Educational Research, 17, 3-18.

Osterlind, S. J. (1989). Constructing test items. Boston, MA: Kluwer.

Pring, R. (1971). Bloom's taxonomy: A philosophical critique. Cambridge Journal of Education, 1, 83-[illegible]

Resnick, L. B. (1987). Education and learning [illegible] k. (Report

No. SP02-96-52). Washington, DC: National Academy. (ERIC Document Reproduction Services No. ED289 832)

Roid, G. H., & Haladyna, T. M. (1982). A technology for test-item writing. New York, NY: Academic Press.

Seddon, G. M. (1978). The properties of Bloom's taxonomy of educational objectives for the cognitive domain. Review of Educational Research, 48, 303-323.

Siegler, R. S., & Shrager, J. (1984). A model of strategy choice. In C. Sophian (Ed.), Origins of cognitive skills (pp. 229-293). Hillsdale, NJ: Erlbaum.

Siegler, R. S. (1988). Individual differences in strategy choices: Good students, not-so-good students, and perfectionist. Child Development, 59, 833-851.

Smith, F. M. (1984). Constructing and using achievement tests in the classroom: A competency based text. New York, NY: Peter Lang.

Snow, R. E. (1978). Theory and method for research on aptitude processes. Intelligence, 2, 225-278.

Snow, R. E., & Peterson, P. L. (1985). Cognitive analyses of tests: Implications for redesign. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics (pp. 149-166). New York: Academic Press.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn

(Ed.), Educational Measurement (3rd ed., pp. 263-331). New York: American Council on Education, Macmillian.

Sockett, H. (1971). Bloom's taxonomy: A philosophical critique. Cambridge Journal of Education, 1, 16-25.

Stahl, R. J., & Murphy, G. T. (1981, April). The domain of cognition: An alternative to Bloom's cognitive domain within the framework of an information processing model. Paper presented at the meeting of the American Educational Research Association, Los Angeles.

Sternberg, R. J., & Weil, E. W. (1980). An aptitude-strategy interaction in linear syllogistic reasoning. Journal of Educational Psychology, 72, 226-234.

Tiemann, P. W., & Markle, S. W. (1973). Re-modeling a model: An elaborated hierarchy of types of learning. Educational Psychologist, 10, 147-158.

Appendix

The Mathematics Achievement Subscale (MAS). Questions for the MAS were adapted directly from Alberta Education Grade 6 Achievement Tests in Mathematics, with permission. The cognitive level and content area for each item are provided in Figure 2. The correct solution is marked with an asterisk. The practice items (i.e., P1 to P3) are also included.

P1. 2 x 3 =

P2.
```
     14
  X  13
  ----------
```

P3. 2763.28 rounded to the nearest whole number is

A. 2763*

B. 2764

C. 2753.0

D. 2763.3

1. 0.0476 rounded to the nearest thousandth is

A. 0.05

B. 0.048*

C. 0.047

D. 0.04

2. Amy earns $24.50 per week. She earns 3.5 times as much as Jacob earns per week.

   How much does JACOB earn per week?

   A. $ 3.50

   B. $ 7.00*

   C. $ 28.00

   D. $ 85.75

3. 3/4 written as a decimal is

   A. 0.25

   B. 0.75*

   C. 3.4

   D. 4.3

4. If 970 is DIVIDED BY 10, and 3 is added to the quotient, the answer is

   A. 9703

   B. 9700

   C. 103

   D. 100*

5. (5 x 1000) + (2 x 10) + (1 x 1) + (9 x 0.1) + (6 x 0.01) written in standard notation is

A. 502.196

B. 521.096

C. 5021.96*

D. 5219.6

6. Mrs. Dawson bought five hockey tickets. Each ticket had a different number. The tickets were numbered in order. The average of the ticket numbers was 24. The numbers of the tickets could have been

A. 24, 24, 24, 24, 24

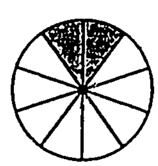
B. 23, 24, 25, 26, 27

C. 22, 23, 24, 25, 26*

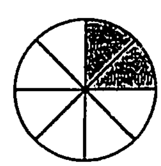
D. 21, 22, 23, 24, 25

7. Which circle has one-fifth of its area shaded?

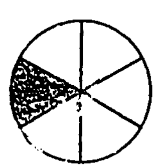
A.*



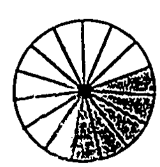
B.



C.



D.



8. The best estimate of 51 x 19 is

A. 10 000

B. 5 000

C. 1 000*

D. 500

9. Find the number.

The hundreds place is one.

The tens place and the ones place are zero.

The tenths place is two times the hundreds place.

The hundredths place is four times the tenths place.

The number is

A. 100.028

B. 100.24

C. 100.26

D. 100.28*

10. Louis wins 6 Star Wars figures for each 5 that Pam wins. If Louise has 24 Star War figures, then Pam has

A. 30

B. 25

C. 20*

D. 18

11. To check the division problem at the right,

A. add 36 to 22, and then multiply the sum by 30

B. add 30 to 22, and then multiply the sum by 36

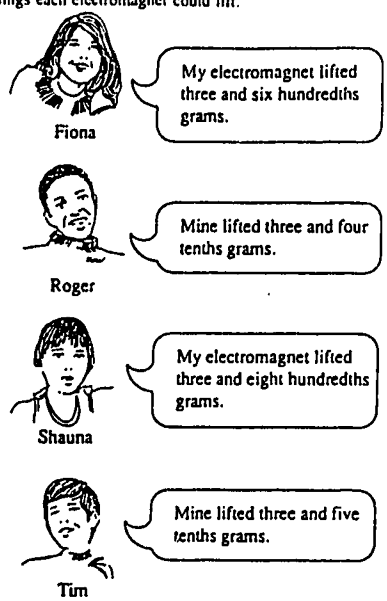
C. multiply 30 by 22, and then add 36 to the product

D. multiply 30 by 36, and then add 22 to the product*

```
     36
30)1102
    90
    202
    180
     22
```

12.

In a science investigation, four students each built an electromagnet. They tested the electromagnets for strength by measuring the mass of iron filings each electromagnet could lift.



Whose electromagnet lifted the greatest mass?

A. Fiona's

B. Roger's

C. Shauna's

D. Tim's*

13. 61 000 - 100 equals

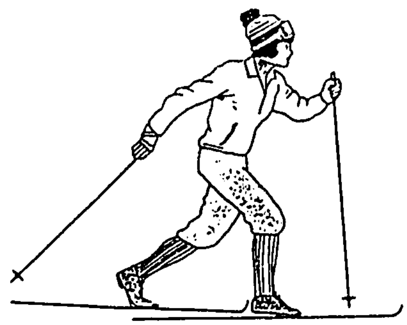
    A. 610*

    B. 6100

    C. 61

    D. 601

14. A cross-country skier goes 3 kilometres in 15 minutes.



At this rate, how long will it take the skier to go 15 km?

    A. 150 min

    C. 90 min

    C. 75 min*

    D. 45 min

15. 90% of 20 is

    A. 9

    B. 17

    C. 18*

    D. 19

16. Which group of integers is arranged from LARGEST to SMALLEST?

    A. -3, 6, 0, 1, 3

    B. -6, -3, 0, 1, 6

    C. 3, 0, -1, -3, -6*

    D. 6, 3, -3, 0, 1

17. There are 80 men on a ship. If they must abandon the ship, what is the SMALLEST number of 7-man lifeboats needed to save everyone?

    A. 17

    B. 12*

    C. 11

    D. 10

18. Fabric for curtains costs \$ 12.50 per metre. If one roll of fabric costs \$ 78.50, how many metres of fabric are there in the roll?

    A. 6.28*

    B. 66

    C. 91

    D. 981.25