A study of possible factors that influence the construction of teacher-made problems that assess higher-order thinking skills
by David Lynn Harpster

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Education in Secondary Curriculum and Instruction
Montana State University

Abstract:
The problem addressed by this study was to analyze statistically the possible influence of several profession-related factors on the levels of questions Montana public high school mathematics teachers would use in their classrooms to assess higher-border thinking skills.

Questionnaires were sent to 220 public high school mathematics teachers in the state of Montana during the winter of 1998. Chi-square tests of independence were used to determine if there was a relationship between the categorization of the teachers' written examples of a question that tests for higher-order thinking skills and o the amount of SIMMS professional development that the teachers had received; o the use of the SIMMS IM curriculum in the teachers' classrooms; o the number of credit hours that the teachers had accumulated from college classes in assessment or evaluation; o the highest academic degree obtained by the teachers; and o the educational objective from Bloom's taxonomy that the teachers ranked most important for their classroom.

ANOVA tests were used to determine if there was a significant difference between o the mean number of hours of professional development for teachers who wrote Level I, Level II, Level III, or Level IV questions as examples of questions that test for higher-order thinking skills; o the mean number of years teaching mathematics for teachers who wrote Level I, Level II, Level III, or Level IV questions as examples of questions that test for higher-order thinking skills.

To help interpret the results of this study, telephone interviews were conducted with ten of the teachers who had returned questionnaires.

No significant relationship or difference was found for any of the factors stated above. Almost 60% of the teachers participating in this study wrote a question that assessed lower-order thinking skills when asked to write a question that assessed higher-order thinking skills.

Professional development, college measurement courses, and continuing education do not seem to impact teachers' assessment of higher-order thinking skills; furthermore, we cannot assume that teaching experience alone, without proper training and classroom support, will improve this situation.

A STUDY OF POSSIBLE FACTORS THAT INFLUENCE THE CONSTRUCTION OF TEACHER-MADE PROBLEMS THAT ASSESS HIGHER-ORDER THINKING SKILLS

by

David Lynn Harpster

A thesis submitted in partial fulfillment

of the requirements for the degree

of

Doctor of Education

in

Secondary Curriculum and Instruction

MONTANA STATE UNIVERSITY- BOZEMAN
Bozeman, Montana

May 1999

APPROVAL

of a thesis submitted by

David Lynn Harpster

This thesis has been read by each member of the thesis committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the College of Graduate Studies.

Dr. William Hall (Co-Chairman) William D. Hall (Signature) 5.6.99 Date

Dr. Maurice Burke (Co-Chairman) Maurice Burke (Signature) May 6, 1999 Date

Approved for the Department of Education

Dr. Gloria Gregg Gloria Gregg (Signature) May 7, 1999 Date

Approved for the College of Graduate Studies

Dr. Bruce R. McLeod Bruce R. McLeod (Signature) 5-8-99 Date

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a doctoral degree at Montana State University-Bozeman, I agree that the Library shall make it available to borrowers under rules of the Library. I further agree that copying of this thesis is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for extensive copying or reproduction of this thesis should be referred to University Microfilms International, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom I have granted the "exclusive right to reproduce and distribute my dissertation in and from microform along with the non-exclusive right to reproduce and distribute my abstract in any format in whole or in part."

Signature David Hayden

Date May 7, 1999

TABLE OF CONTENTS

TABLE OF CONTENTS - Continued

LIST OF TABLES

ABSTRACT

The problem addressed by this study was to analyze statistically the possible influence of several profession-related factors on the levels of questions Montana public high school mathematics teachers would use in their classrooms to assess higher-order thinking skills.

Questionnaires were sent to 220 public high school mathematics teachers in the state of Montana during the winter of 1998. Chi-square tests of independence were used to determine if there was a relationship between the categorization of the teachers' written examples of a question that tests for higher-order thinking skills and

- the amount of SIMMS professional development that the teachers had received;
- the use of the SIMMS IM curriculum in the teachers' classrooms;
- the number of credit hours that the teachers had accumulated from college classes in assessment or evaluation;
- the highest academic degree obtained by the teachers; and
- the educational objective from Bloom's taxonomy that the teachers ranked most important for their classroom.

ANOVA tests were used to determine if there was a significant difference between

- the mean number of hours of professional development for teachers who wrote Level I, Level II, Level III, or Level IV questions as examples of questions that test for higher-order thinking skills;
- the mean number of years teaching mathematics for teachers who wrote Level I, Level II, Level III, or Level IV questions as examples of questions that test for higher-order thinking skills.

To help interpret the results of this study, telephone interviews were conducted with ten of the teachers who had returned questionnaires.

No significant relationship or difference was found for any of the factors stated above. Almost 60% of the teachers participating in this study wrote a question that assessed lower-order thinking skills when asked to write a question that assessed higher-order thinking skills.

Professional development, college measurement courses, and continuing education do not seem to impact teachers' assessment of higher-order thinking skills; furthermore, we cannot assume that teaching experience alone, without proper training and classroom support, will improve this situation.

CHAPTER 1

PROBLEM STATEMENT AND REVIEW OF LITERATURE

Introduction

The Impact of Assessment

Assessment substantially impacts the public school classroom. Crooks (1988), in his review of the literature on classroom assessment, found that assessment affects students in several ways:

> For instance, it [assessment] guides their judgment of what is important to learn, affects their motivation and self-perceptions of competence, structures their approaches to and timing of personal study (e.g., spaced practice), consolidates learning, and affects the development of enduring learning strategies and skills. It appears to be one of the most potent forces influencing education. (p. 467)

That assessment is such a potent force is not unexpected considering that, on average, formal written tests alone occupy from five to fifteen percent of a student's time (Crooks, 1988). The lower figure is more typical of elementary school students whereas the higher figure is more typical of high school students. In addition, teachers may spend as much as 20 to 30% of their professional time directly involved in assessment-related activities. (Stiggins, 1988).

Higher-Order Thinking Skills

Assuming that assessment substantially impacts the public school classroom, then what thinking skills should be emphasized through assessment? The demands of business and industry suggest that higher-order thinking skills should be emphasized in education. For example, Cross (1985) believes that both small businesses and industry will need workers who can synthesis and utilize information; higher-order thinking skills according to Bloom's taxonomy. Likewise, Daggett (1996) recommends an emphasis on the transfer of knowledge to unpredictable situations. He believes that this is an area where American students lag behind both European and Asian students, with possibly profound economic and social consequences in the future.

Research on the levels (according to Bloom's taxonomy) at which teachers assess their students indicates that, generally, they assess lower-order thinking skills. For instance, Fleming and Chambers (1983) in their research of the Cleveland school district concluded that ". . . teacher-made tests do not require students to display higher order abilities". (p. 30) Nine years later, in his study of Georgia secondary mathematics teachers, Cooney (1992) found that ". . . many teachers equate a deeper understanding of mathematics with the ability to perform more difficult computations or to solve one step problems". (p. 18)

The SIMMS Project

In 1991, the state of Montana received funding from the National Science Foundation (NSF) for the Systemic Initiative for Montana Mathematics and Science (SIMMS) Project. The SIMMS Project had nine objectives:

1. Promote integration in science and mathematics education.
2. Redesign the 9-12 mathematics curriculum using an integrated interdisciplinary approach for all students.
3. *Develop and publish curriculum and assessment materials for grades 9-16* [italics added].
4. Incorporate the use of technology in all facets and at all levels of mathematics education.
5. Develop an action plan to increase the participation of females and Native Americans in mathematics and science.
6. Establish new certification standards and recertification standards for teachers.
7. Redesign teacher preparation programs using an integrated interdisciplinary approach.
8. *Develop an extensive inservice* [sic] *program in mathematics grades 9-16 to prepare teachers for integrated programs* [italics added].
9. Develop the support structure for legislative action, public information and general education of the populace. (SIMMS Project, Summer, 1997, pp. 2-7)

There has to date, been no research on the impact of SIMMS professional development on the assessment practices of SIMMS certified teachers.

Statement of the Problem

The problem addressed by this study was to statistically analyze the relationship between specific factors (listed below) and the levels of questions Montana public high school mathematics teachers would use in their classroom to assess higher-order thinking skills. This study was conducted to answer the following questions. Using Cooney's taxonomy for categorizing mathematical test questions,

1. What levels of questions do Montana public high school mathematics teachers write to assess the higher-order thinking skills of their students?
2. Is there a significant difference in the amount of professional development that Montana public high school mathematics teachers have received based upon

the categorization of their written examples of questions that they would use in their classroom to test for higher-order thinking skills?

3. Is there a significant difference in the number of years teaching mathematics for Montana public high school mathematics teachers based upon the categorization of their written examples of questions that they would use in their classroom to test for higher-order thinking skills?
4. Is there a pattern or relationship between the categorization of Montana public high school mathematics teachers' written examples of questions they would use in their classroom to test for higher-order thinking skills and
   - the amount of SIMMS professional development that they have received;
   - their use or non-use of the SIMMS IM curriculum;
   - the number of credit hours that they have accumulated from college classes in assessment or evaluation;
   - their highest academic degree;
   - the order in which they rank the importance of the educational objects from Bloom's taxonomy.

Furthermore, what rationale do teachers give for determining if a question assesses higher-order thinking skills?

The Importance of the Study

There is ample evidence from the literature suggesting that assessment, through teacher-generated tests, has an impact upon what and how students learn. In addition, studies of teacher-made tests indicate that teachers often assess predominantly lower-order thinking skills, relative to Bloom's taxonomy. What the literature does not tell us is the cognitive levels of the questions that teachers would use in their classroom to test for higher-order thinking skills and what factors may influence this. The study by Cooney (1992) was the first step in providing answers to these questions, with respect to secondary mathematics teachers. However, in the absence of further research to support or confirm his conclusions, we cannot generalize his results across the whole population of public high school mathematics teachers in this country.

The importance of this study is that it informs mathematics educators of the cognitive levels, according to Cooney's taxonomy, of questions that a specific population of teachers (Montana public high school mathematics teachers) would use in their classroom to assess higher-order thinking skills. It also sheds some light on the impact that SIMMS professional development, the SIMMS IM curriculum, and other factors have had upon the cognitive levels of questions that teachers would use in their classroom to assess higher order thinking skills.

Before we as mathematics educators suggest methods to remedy the problem of public high school mathematics teachers assessing, predominantly, lower-order thinking skills, we must first know what levels of questions teachers believe assess higher-order thinking skills and what factors may influence those beliefs. If we do not determine the specific factors that influence teacher beliefs about the assessment of higher-order thinking skills, then we may be throwing solutions at problems that we do not adequately understand.

Definition of Terms

For the purpose of this study, the researcher will use the following definitions:

Assessment: The comprehensive accounting of an individual's or group's functioning within a specific subject area or application of that subject area. (Adapted from Webb, 1992, pp. 662-663)

Cooney's Taxonomy: A modification of Bloom's taxonomy used to classify questions from tests and quizzes in mathematics. Cooney's taxonomy consists of the following levels:

Level 1 questions are those that require recognition on the part of the student or a simple computation. This category is synonymous with *Knowledge* in Bloom's taxonomy.

Level 2 questions require a student to make some decision, but once made the solution process is straightforward (e.g., a simple one-step problem). This category is synonymous with *Comprehension* in Bloom's taxonomy.

Level 3 questions are application questions or multiple-step problems in which a student must make several decisions about the solution process or what operations to use. This category is synonymous with *Application* in Bloom's taxonomy.

Level 4 questions are non-routine or open-ended problems. This category is synonymous with the combination of *Analysis*, *Synthesis*, and *Evaluation* from Bloom's taxonomy.

Evaluation: Assigning value to the results of an assessment.

Higher-Order Thinking Skills: These would include the following skills from Bloom's taxonomy: *Application*, *Analysis*, *Synthesis*, and *Evaluation* (Levels 3, 4, 5, and 6, respectively).

High School: Grades nine through twelve in the public schools.

Lower-Order Thinking Skills: These would include the following skills from Bloom's taxonomy: *Knowledge* and *Comprehension* (Levels 1 and 2, respectively).

Non-SIMMS IM: A Montana teacher who has received less than three hours of SIMMS professional development.

Open-ended Question: Using Frederiksen's (1994, p.538) definition, a question is open-ended if it (a) is non-algorithmic; (b) is complex; (c) may yield more than one solution, each of which has different benefits and/or disadvantages; (d) may require nuanced judgments -- sensitivity to subtle aspects of a problem situation; (e) may involve the application of multiple and sometimes conflicting criteria; (f) may involve uncertainty because all of the facts are not known; (g) may involve self-regulation of the thinking process (metacognition); (h) may require one to find structure in apparent disorder; and (i) requires effort because of the complexity of problems.

SIMMS IM certified: A Montana teacher who has received a minimum of two weeks (60 or more hours) of SIMMS professional development.

SIMMS IM trained: A Montana teacher who has received from one to two weeks (30 hours to 59 hours) of SIMMS professional development.

SIMMS IM introduced: A Montana teacher who has received from one-half of a day to less than one week (3 hours to 29 hours) of SIMMS professional development.

Substantial Native American Population: Thirty-three percent or more of the students are Native American according to data collected by the Office of Public Instruction in the state of Montana.

Review of the Literature

Assessment

In 1988, Crooks published a review of research literature related to assessment practices. He found that assessment affects students by

- reactivating or consolidating prerequisite skills or knowledge prior to introducing new material;
- focusing attention on important aspects of the subject;
- encouraging active learning strategies;
- giving students opportunities to practice skills and consolidate learning;
- providing knowledge of results and corrective feedback;
- helping students to monitor their own progress and develop skills of self evaluation;
- guiding the choice of further instructional or learning activities to increase mastery;
- helping students feel a sense of accomplishment. (p. 443)

Frederiksen (1984) observed that "tests tend to increase the time and effort spent in learning and teaching what the tests measure and . . . decrease efforts to learn and teach skills not measured by the test." (p. 193) Further, he notes that "If educational tests fail to represent the spectrum of knowledge and skills that ought to be taught, they may introduce bias against teaching important skills that are not measured." (p. 193)

It is important to focus on the assessment practices of high school mathematics teachers because research indicates that this group of teachers tends to spend more time assessing students through teacher-made tests than through any other method of assessment. In addition, these tests have a large influence on how high school mathematics teachers evaluate students when determining report card grades. In a nationwide study of teachers, including 363 high school teachers, Herman and Dorr-Bremme (1984) found that on a scale from 1 to 4, where 1 indicates unimportant and 4 indicates crucial importance, secondary school teachers rated teacher-made test at 3.65 in terms of deciding on report card grades. The authors also found that about 76% of the time tenth-grade mathematics teachers devoted to testing was apportioned to administering teacher-made tests. The importance of teacher-made tests to high school teachers was confirmed by Stiggins and Bridgeford (1985) in a nationwide study of 228 teachers in public schools. They concluded that "the higher the grade level, the greater the tendency for teachers to report using their own assessments rather than published tests." (p. 281)

In 1992, Marso and Pigge published a summary of their review of approximately 225 studies addressing K-12 classroom teachers' knowledge of skills related to the development and use of teacher-made tests. With regard to teachers' testing practices, attitudes, and beliefs, the researchers arrived at the following conclusions:

- Teachers rely on teacher-made tests to a much greater extent than standardized tests and district state competency tests for making decisions about individual pupils.
- Teachers believe that self-constructed assessments generally better meet the instructional needs of their classes than do assessments derived from other sources such as workbooks or textbooks.
- Teachers believe that they are less proficient in testing skills when compared to their proficiencies in other professional skill areas.
- Teachers generally report that they have deficiencies in testing and measurement, feel that their self-constructed tests could be improved, and would like inservice [sic] training in tests and measurements if this training were oriented toward practical classroom needs, but they tend to be confident about their general abilities and knowledge.
- Teachers believe that essay tests as compared to objective tests are impractical and disliked by pupils but result in greater study efforts and usually measure at higher cognitive levels.
- Teachers believe that testing, evaluation, and grading activities are among their more demanding and less pleasant classroom responsibilities.
- For most preservice [sic] and inservice [sic] teachers it appears that their knowledge of classroom testing practices and principles is inadequate to meet classroom evaluation needs, and it appears that little progress has been made in overcoming this inadequacy during the past quarter century.
- Neither inservice [sic] training, if provided, nor increased years of teaching experience appear to improve either classroom teachers' testing knowledge or their test construction skills as measured by paper and pencil tests and as revealed by direct analysis of construction flaws found on their self-constructed tests.
- Teachers appear to value the importance of having higher cognitive functioning questions on teacher-made tests, but they infrequently use such questions; they tend to over-estimate the number of higher order questions used on their tests; and they have difficulty identifying and writing test questions that function beyond the knowledge level. (pp. 23-27)

Another study that relates to teachers' assessment practices was conducted by Newman and Stallings (1982). In May of 1980, Newman and Stallings mailed out 1500 questionnaires to K-12 teachers in Atlanta, Georgia, Pensacola, Florida, and Mobile, Alabama for the purposes of determining how well teachers understand classroom testing principles and gain information on the measurement preparation and classroom practices of teachers. The return rate was approximately 21.5%. The researchers found that "teachers use teacher-made tests extensively and spend a substantial amount of professional time constructing, administering, scoring, and interpreting them." (p. 8) Furthermore, it was concluded that although the data suggest a trend towards more measurement work in teacher training programs, as compared to a study conducted in 1967, this did not translate into a better understanding of classroom testing principles.

Higher-Order Thinking Skills

Teaching and assessing higher-order thinking is considered a priority by many authors. Cross (1985) believes that the advent of the "information explosion" will require business and industry in the 21st century to place a premium on employees who can synthesize large amounts of information. Daggett (1994) notes that in the 1950's, about 60% of the jobs in the United States were unskilled, whereas recently that percentage has dropped to 35%. By the year 2000 he predicts that only 15% of the jobs in the United States will be unskilled. The skill that Daggett affirms will be in high demand in the future is the ability to transfer knowledge to new situations. He attributes this to the demand of small businesses for workers who can solve problems without procedures to follow or managers to assist them. Reviewing statistics released by the U. S. Department of Labor in November 1995, Daggett (1996) found that the American work force employed by companies with 20 or fewer employees is increasing at an annual rate of 7.5%.

Many educators contend that higher-order thinking skills should be taught and assessed in the public schools. Jones (1988), in his review of the literature on assessment, states that the Alexander-James Study Group recommended that mathematics and science teachers assess problem-solving and higher-order thinking skills so as to "'allow . . . inferences about the thought processes contributing to the answer' (Alexander & James, 1987, p. 23)." (p. 235) Lohman (1993) stresses that "transfer is an important, if not the most important goal of education" and concludes that "Consequently, one of the most important functions of tests should be to estimate how far students can transfer their learning." (p. 48) Frederiksen (1994) cites six properties that tests of the future should have: Two of them that pertain to the present discussion are

- It would be ideal if the new tests would collectively assess all of the important aspects of learning, not just factual knowledge. This means that tests should assess problem solving not only in the conventional sense, but also with the relevant underlying skills and abilities that are involved.
- The new tests should encourage generalization of learning. The use of a wide variety of problems in different situations and settings should facilitate transfer to new situations. (p. 534)

Similar statements about the need to assess higher-order thinking are often found in the literature. See, for example (Research into Practice Project, 1984), (Peterson, 1988), or (National Academy of Sciences, 1991).

Stiggins, Rubel, and Quellmalz (1985) wrote, "We believe that the new wave of educational improvement is focusing on the improvement of reasoning skills. These are the new basic skills." (p. 37) They continue by stating that "One key to the success of these efforts will be each individual teacher's ability to measure thinking skills in a valid and reliable manner." (p. 37)

An interesting study that relates instructional goals to the emphasis, by teachers, of higher-order thinking was conducted by Raudenbush, Rowan, and Cheong (1993).

Questionnaires were mailed to all teachers in 16 high schools in the states of California and Michigan to explore variations in emphasis on teaching higher-order thinking in secondary schools. The return rate varied from 50 to 100% with a median of 75%. The researchers concluded that, with respect to emphasis on higher-order thinking, "differentiated instructional objectives are strongly institutionalized in the secondary mathematics and science curricula" (p548) and that there are institutional obstacles to promoting higher-order thinking in secondary mathematics and science.

Assessment of Higher-Order Thinking Skills

Several studies between 1983 and 1997 have documented the lack of assessment of higher-order thinking skills through testing in the public schools. A court order for desegregation of the Cleveland school district required a systematic technical analysis of teacher-made tests in the district to determine if they were being developed, administered, scored, and used in a nondiscriminatory manner. Fleming and Chambers (1983) reported that the review process generated information about the technical aspects of 342 (8800 test questions) teacher-made tests in four subject areas (including mathematics), grades 1 through 12, over a two-year period. A second review process classified test questions according to behavioral categories delineated by Bloom (1981): *knowledge of terms*; *knowledge of facts*; *knowledge of rules and principles*; *skill in using processes and procedures*; *ability to make translations*; and *ability to make applications*. Eight observations about how teachers test in the Cleveland school district emerged:

1. Teachers use short-answer questions most frequently in their test making.
2. Teachers, even English teachers, generally avoid essay questions.
3. Teachers use more matching items than multiple-choice or true-false items.

4. Teachers devise more test questions to sample *knowledge of facts* than any of the other behavioral categories studied.
5. When categories related to *knowledge of terms*, *knowledge of facts*, and *knowledge of rules and principles* are combined, almost 80% of the test questions reviewed focus on these areas.
6. Teachers develop few questions to test behaviors that can be classified as *ability to make applications*.
7. Comparison across school levels shows that junior high school teachers use more questions to tap *knowledge of terms*, *knowledge of facts*, and *knowledge of rules and principles* than elementary or senior high school teachers (94%, 69%, 69% respectively).
8. At all grade levels, teacher-made mathematics and science tests reflect a diversity of behavioral categories, since they typically feature questions in all six behavioral categories. (p. 32)

Over 73% of the mathematics questions surveyed tested at the levels *knowledge of terms*, *knowledge of facts*, *knowledge of rules and principles*, and *skill in using processes and procedures*. *Ability to make translations* and *ability to make applications* were evaluated by less than 27% of the questions.

Support for the generalization of Fleming and Chambers (1993) observations across not only the nation, but also across subject areas, is found in a study by Carter (1984). This study was conducted using a sample of 310 secondary reading and language teachers from a four-state area in the south central region of the United States. Teachers were given ten multiple-choice items and asked to identify the reading skills that they believed the items were designed to test. They were then given an objective specifying one of four reading skills (main idea, detail, inference, and prediction) for which they were to develop a multiple choice item that assessed for the achievement of that objective. Results of the study indicate that many of the teachers were unable to recognize the particular skill being tested by individual test items. Also, "Teachers spent more time on and had more difficulty in developing items tapping higher-level skills than writing items to test lower level cognitive skills." (p. 59)

A study by Ball, Doss, and Dewalt (1986) of 74 social studies teachers in two southeastern states was conducted through a questionnaire and the collection of classroom tests. It was found that "Slightly less than half of the teachers (47%) indicated that higher level objectives were important in their students' learning and fewer (24%) said they predominantly used higher level objectives in their classroom teaching or indicated they used them when constructing classroom tests (26%)." (p. 29) It was found that the vast majority of test questions on classroom tests (from 95.6 to 99.7%) assessed lower-level objectives.

Utilizing interviews, observations, and samples of teachers' assessment instruments, Stiggins, Griswold, and Wikelund (1989) gathered data on the assessment practices of 36 volunteer teachers (grades 2-12) in a suburban school district in the northwestern region of the United States. The study focused on four content areas: mathematics, science, social studies, and language arts. Besides other methods of data collection, the researchers asked the teachers to provide four to six samples of paper and pencil assessments recently used in their classroom. It was determined that although the vast majority of teachers had been trained to teach higher-order thinking skills, fewer than one-third had participated in more than one training session in the assessment of higher-order thinking skills and one-third had not received any training at all in this area. Training in the assessment of higher-order thinking skills was found to be more common among teachers at the middle and high school level. The largest percentage of test items collected by the investigators (approximately 46%) tested for recall of facts and information; *Knowledge* and *Comprehension* according to Bloom's taxonomy. This pattern was consistent across grade levels. If we focus only upon the mathematics test items we find that 19% of the items tested for recall (*Knowledge* according to Bloom's taxonomy), nine percent tested for comparison (*Analysis* according to Bloom's taxonomy) and all of the rest (72%) tested for inference (*Application* and *Synthesis* according to Bloom's taxonomy). The later results (categorization of mathematics test items) seem inconsistent with other research. This may be explained by the observation that the researchers considered a problem to be higher-order if a student was expected to solve a problem without being prompted by the teacher for the required steps needed to solve the problem. The researchers considered professional development that focuses on the assessment of higher-order thinking skills, a critical factor in changing teachers' assessment practices.

A study by Cooney (1992) was conducted to provide a basis for understanding the nature of secondary mathematics teachers' (grade 7-12) evaluation practices. The study consisted of three phases: the administration of a questionnaire aimed at obtaining general information about teachers' assessment practices, a second questionnaire that focused on soliciting teachers' reactions to five non-traditional assessment items, and an interview with selected teachers. Copies of the Phase I questionnaire were sent to all instructors in the State of Georgia who were teaching Eisenhower Plan supported mathematics or mathematics education courses for teachers. Two hundred and seventy-nine teachers responded to the Phase I questionnaire and of these, 201 teachers taught mathematics in grades 7-12. Phase II questionnaires were sent to the 201 teachers who had completed the first questionnaire and taught mathematics in grades 7-12. Of those, 102 completed the questionnaire. Twenty of the 102 teachers completing the second questionnaire were selected to be interviewed at their schools, but only 18 could participate. The teachers who were selected to be interviewed reflected a wide range of evaluation practices, varied responses to the Phase II questionnaire, and different geographical and school settings.

The Phase I questionnaire was designed to provide descriptive information about the contexts in which the teachers taught the previous school year (1989-1990), the nature of their assessment practices, the importance attributed to different purposes for assessment, and their perception of how math is learned. As part of the first questionnaire, teachers were asked to write or draw a typical problem that they had given their students that they believed tested for a minimal understanding of the topic. Similarly, they were asked to do the same thing for a typical problem that they believed tested for a deep and thorough understanding of the topic. Their responses were categorized using the following four levels:

1. Simple computation or recognition;
2. Comprehension. Students must make some decision but once made the solution process is straightforward, e.g., a simple one step problem;
3. An application or multistep problem in which [the] student must make several decisions about [the] solution process or what operations to use; and
4. Nonroutine or open-ended problem. (Cooney, 1992, p. 7)

In addition, teachers were asked to give a rating ranging from 1 (very important) to 5 (of no importance) to the following purposes of assessment:

1. Certify level of mathematical competence
2. Identify students' misconceptions
3. Provide feedback to students on [their] progress

4. Help students organize knowledge
5. Increase students' motivation to learn
6. Assist teachers in keeping students on task. (p. 10)

The Phase II questionnaire presented five non-traditional items and asked the teachers to construct an ideal response, identify the mathematical content and processes each item assessed, indicate the likelihood that they would use the items with their students, and describe the type of student or class for which they would most be inclined to use the items. The interview protocol was developed to obtain more detailed information about the teachers' responses to the first two questionnaires.

It was found that only 2 of the 18 teachers interviewed had any substantial knowledge of the NCTM Curriculum and Evaluation Standards for School Mathematics (1989) (henceforth referred to as the Standards) and none of the teachers had any knowledge of the Evaluation section of the Standards. Cooney suggested that the Standards be used as a leverage point for teacher education programs. There is evidence from this study that many teachers are assessing students on a narrow range of outcomes such as computations and one-step problems. A high proportion (57%) of the teachers in the study equated a deeper understanding of mathematics with the ability to perform difficult computations or to solve one-step problems. Less experienced teachers were less likely to write Level 3 or Level 4 items to test for a deep and thorough understanding than their experienced counterparts. The author suggested that beginning teachers receive some sort of pre-service or in-service training that focuses on generating such questions. Teachers liked the non-traditional test items in the second questionnaire that required the generation of a specific number more than the test items that required an explanation or the generation of a counterexample. They rationalized this preference by expressing concern about the time that it would take to grade more open-ended questions. Other frequently mentioned concerns were

1. Students' ability to read and write are [sic] too low to profit from responding to such items.
2. Students have poor thinking skills.
3. Students have short attention spans. (Cooney, 1992, p. 17)

Teachers indicated that the most important purposes of assessment are to certify competence, identify misconceptions so that instruction can be adjusted, and to provide feedback to students on their progress.

Cooney concluded that skill performance, rather than conceptual understanding, is what many teachers associate with mathematical competence and that reform in the mathematics classroom is not a widespread phenomena and is not likely to become one unless significant in-service programs are provided. He asserted that teachers need to see mathematics as more than just a sequence of well-defined steps. A final observation made by the author was that we cannot expect reform in mathematics education to occur in the absence of assessment methods that reflect a broader view of mathematics. Students and teachers must see mathematics as a science of reasoning, posing and solving problems, and as a means of communication.

The most recent study of assessment practices by high school mathematics teachers was conducted by Senk, Beckmann, and Thompson (1997). Nineteen classes in five high schools in three Midwestern cities were studied utilizing data obtained from questionnaires and the collection of all written assessment instruments used in each classroom. It was determined that "the most frequently used assessment tools were tests and quizzes, with these determining about 77% of students' grades." (p. 187) In common with the majority of the other studies of classroom tests mentioned above, tests items, generally, "were low level, were stated without reference to a realistic context, involved very little reasoning, and were almost never open-ended." (p. 187) It is interesting to note that the "teachers' knowledge and beliefs, as well as the content and textbook of the course, influenced the characteristics of test items and other assessment instruments." (p. 187)

Professional Development

In each of the studies reviewed in the previous section, the authors suggested that teachers receive further training in assessment either through pre-service instruction or professional development. Fielding and Shaughnessy (1990) conducted a study of the impact of professional development using teachers from ten high schools, two from each of five school districts in western Oregon. The goal of their program was to enhance high school science teachers' skills in designing and using tests as aids to instruction. Teachers from five of the schools participated in school-based workshops whereas teachers from the other five schools received no training. Based upon informal comments and written responses from teachers who participated in the workshops, the workshops were considered a success. A subsequent impact analysis utilizing student and teacher interviews and classroom observations revealed that there were no clear differences between teachers whom had received training and those who had not. This was attributed, initially, to the limited training time and follow-up support. However, the researchers did not consider this an adequate explanation. They concluded that the demands of classroom management, the curriculum's overemphasis on content coverage, and a lack of organizational commitment on the part of the schools, rather than inadequacies in the training program, resulted in their training program having little or no lasting impact upon the assessment practices of teachers in the experimental group.

In 1989, Sparks and Loucks-Horsley organized what was then known about professional development into five models:

1. Individually-Guided Staff Development
2. Observation/Assessment
3. Involvement in a Development/Improvement process
4. Training
5. Inquiry

The SIMMS Project introduced Montana mathematics teachers to the SIMMS IM curriculum by utilizing three of those models: Involvement in a Development/Improvement Process, Training, and Observation/Assessment. The SIMMS Project, over a four-year period (1992-1996), employed approximately 70 teachers to help develop the SIMMS IM curriculum. In addition, about 80% of all high school mathematics teachers in the state of Montana have received some degree of SIMMS IM professional development through workshops, seminars, summer institutes, extended studies courses, and in-service programs. (SIMMS Project, Summer, 1997, p. 10) Initially, when the curriculum was introduced into high school classrooms, site visits were used to provide feedback to participating teachers. Unlike the professional development program utilized by Fielding and Shaughnessy, the SIMMS Project has provided several Montana teachers with extensive professional development related to the SIMMS IM curriculum.

Theoretical Framework

Studies conducted by Fleming and Chambers (1983); Carter (1984); Ball, Doss, and Dewalt (1986); Stiggins, Griswold, Green, and Associates (1987); Cooney (1992); and Senk, Beckmann, and Thompson (1997) indicate that, generally speaking, higher-order thinking skills are not being assessed by teacher-made tests in the public schools. With the exception of both Carter and Cooney, all of those studies focused on the tests and quizzes that teachers used to assess their students. Teachers were not asked to specifically generate a question that assessed higher-order thinking skills. Hence, we cannot determine if teachers have the ability to write questions for their students that assess higher-order thinking skills. In addition, with the exception of the study by Cooney (1992), the research literature does not indicate what factors may affect the levels at which public high school mathematics teachers assess their students. Cooney categorized teacher-made questions according to a taxonomy, created specifically for the study, based upon Bloom's taxonomy and related those results to the teachers' teaching experience.

Bloom (1956) established a theoretical framework for categorizing assessment questions according to behaviors that students were asked to demonstrate. Cooney modified this taxonomy for use in categorizing teacher-made mathematical problems and related those categories to teachers' teaching experience to determine if there was a pattern or relationship. The theoretical framework of this study was based upon this aspect of Cooney's research. This study utilized research methods similar to Cooney's (questionnaires and interviews) with two substantial differences:

1. Statistical comparison and test of relationship were used to analyze the data.
2. Other factors besides teaching experience were tested for relationship to the levels of the problems that teachers wrote.

This study is significant in that it focuses on factors that are directly or indirectly affected by mathematics educators and utilizes statistical techniques to determine if those factors relate to the assessment of higher-order thinking skills by high school mathematics teachers in the state of Montana. This information is necessary for determining the type and content of classes that should be used to prepare high school mathematics teachers so that they have available working-knowledge of how to assess higher-order thinking skills in their classroom.

CHAPTER 2

METHODOLOGY

Quantitative Research Methodology

Population Description and Sampling Procedure

The population for this study was all full-time mathematics teachers in the state of Montana who teach grades nine through twelve in the public high schools. It was originally estimated, based upon the SIMMS Integrated Mathematics Implementation Schools Directory (1997-1998), that there were approximately 576 teachers in the state of Montana teaching mathematics in grades nine through twelve in the public high schools during the 1997-1998 school year and that 30 to 35% of those teachers used SIMMS IM curriculum materials in their classrooms.

Base upon those estimates, 234 teachers from the population described above were selected to participate in this study using a stratified random sample. The sample size was determined by calculating the size of a sample required for a 95 percent degree of accuracy using the formula published by Krejcie and Morgan (1970, p. 607). Midway through the process of collecting the questionnaires, it was determined that the estimate of 576 secondary high school mathematics teachers in the state of Montana was too high since several of the teachers who responded did not meet the criteria of the study. A better estimate for the population of the study would be somewhere between 475 and 500 teachers with a sample size of between 215 to 220 teachers to obtain a 95% degree of accuracy. A simple computation, based upon returned questionnaires, indicated that approximately 35 of the teachers receiving the questionnaire did not meet the criteria of the study so it was decided to add another group of 21 randomly selected teachers to the original sample.

Seven categories of teachers were randomly sampled in proportion to the ratio of the number of teachers in each category to the total population of teachers being sampled. Those categories, as determined by the Office of Public Instruction in the state of Montana, were:

1. Teachers who teach at a Class C high school with a substantial Native American population (see p. 8 for the definition of Substantial Native American Population)
2. Teachers who teach at a Class C high school that does not have a substantial Native American population
3. Teachers who teach at a Class B high school with a substantial Native American population
4. Teachers who teach at a Class B high school that does not have a substantial Native American population
5. Teachers who teach at a Class A high school with a substantial Native American population
6. Teachers who teach at a Class A high school that does not have a substantial Native American population

7. Teachers who teach at a Class AA high school that does not have a substantial Native American population

There are, presently, no Class AA public high schools with a substantial Native American population in the state of Montana.

Null Hypotheses to be Tested

It should be noted that any reference to the categorization of teacher's written questions, for the purpose of this study, is in terms of Cooney's taxonomy. In addition to determining the number and percent of public high school mathematics teachers in the state of Montana who wrote Level 1, Level 2, Level 3, and Level 4 questions, this study tested the following null hypotheses:

1. There is no significant difference between the mean number of hours of professional development for teachers who wrote Level I, Level II, Level III, or Level IV questions as examples of questions that test for higher-order thinking skills.
2. There is no significant difference between the mean number of years teaching mathematics for teachers who wrote Level I, Level II, Level III, or Level IV questions as examples of questions that test for higher-order thinking skills.
3. There is no relationship between the categorization of teachers' written examples of questions that test for higher-order thinking skills and the amount of SIMMS professional development that they have received.
4. There is no relationship between the categorization of teachers' written examples of questions that test for higher-order thinking skills and the use of the SIMMS IM curriculum in their classroom.

5. There is no relationship between the categorization of teachers' written examples of questions that test for higher-order thinking skills and the number of credit hours that they have accumulated from college classes in assessment or evaluation.
6. There is no relationship between the categorization of teachers' written examples of questions that test for higher-order thinking skills and the highest academic degree they have obtained.
7. There is no relationship between the categorization of teachers' written examples of questions that test for higher-order thinking skills and which educational objective from Bloom's taxonomy they rank most important in their classroom.

Categorization of Variables

Categories for Level 1, Level 2, Level 3, and Level 4 questions were sometimes combined when utilizing a chi-square test of independence so that there would be no zeroes in the matrix. This was done by combining Levels 1 and 2 (lower-order objectives) and Levels 3 and 4 (higher-order objectives), respectively.

The variables for Null Hypothesis 1 were *categorization of the teacher's question* and *amount of professional development*. The former variable was categorized in the following way:

1. Teachers who wrote a Level 1 question as an example of a question that assesses higher-order thinking skills
2. Teachers who wrote a Level 2 question as an example of a question that assesses higher-order thinking skills

3. Teachers who wrote a Level 3 question as an example of a question that assesses higher-order thinking skills
4. Teachers who wrote a Level 4 question as an example of a question that assesses higher-order thinking skills

The variables for Null Hypothesis 2 were *categorization of the teacher's question* and *years of experience teaching mathematics*. The former variable was categorized in the following way:

1. Teachers who wrote a Level 1 question as an example of a question that assesses higher-order thinking skills
2. Teachers who wrote a Level 2 question as an example of a question that assesses higher-order thinking skills
3. Teachers who wrote a Level 3 question as an example of a question that assesses higher-order thinking skills
4. Teachers who wrote a Level 4 question as an example of a question that assesses higher-order thinking skills

The variables for Null Hypothesis 3 were *categorization of the teacher's question* and *amount of SIMMS professional development*. The former variable was categorized in the following way:

1. Teachers who wrote a Level 1 question as an example of a question that assesses higher-order thinking skills
2. Teachers who wrote a Level 2 question as an example of a question that assesses higher-order thinking skills

3. Teachers who wrote a Level 3 question as an example of a question that assesses higher-order thinking skills
4. Teachers who wrote a Level 4 question as an example of a question that assesses higher-order thinking skills

The latter variable was categorized in the following way:

1. Teachers who are either SIMMS IM introduced or Non-SIMMS IM
2. Teachers who are either SIMMS IM certified or SIMMS IM trained

The variables for Null Hypothesis 4 were *categorization of the teacher's question* and *use of the SIMMS IM curriculum in the classroom.* The former variable was categorized in the following way:

1. Teachers who wrote a Level 1 or Level 2 question as an example of a question that assesses higher-order thinking skills
2. Teachers who wrote a Level 3 or Level 4 question as an example of a question that assesses higher-order thinking skills

The latter variable was categorized in the following way:

1. Teachers who use SIMMS IM curriculum materials exclusively in one or more of their classrooms and are either SIMMS IM certified or SIMMS IM trained
2. Teachers who use SIMMS IM curriculum materials exclusively in one or more of their classrooms and are either SIMMS IM introduced or Non-SIMMS IM, teachers who use SIMMS IM curriculum materials in their classrooms in conjunction with another textbook or curriculum, or teachers who do not use

SIMMS curriculum materials in their classroom and are either SIMMS IM certified or SIMMS IM trained

3. Teachers who do not use SIMMS curriculum materials in their classrooms and are either SIMMS IM introduced or Non-SIMMS IM

By splitting teachers into these three categories the influence of SIMMS professional development on those teachers who do not teach using the SIMMS IM curriculum was controlled.

The variables for Null Hypothesis 5 were *categorization of the teacher's question* and *number of credit hours*. The former variable was categorized in the following way:

1. Teachers who wrote a Level 1 or Level 2 question as an example of a question that assesses higher-order thinking skills
2. Teachers who wrote a Level 3 or Level 4 question as an example of a question that assesses higher-order thinking skills

The latter variable was categorized in the following way:

1. Teachers who have accumulated zero to three credit hours for college classes in assessment or evaluation
2. Teachers who have accumulated from four to six credit hours for college classes in assessment or evaluation
3. Teachers who have accumulated seven or more credit hours for college classes in assessment or evaluation

The variables for Null Hypothesis 6 were *categorization of the teacher's question* and *highest academic degree obtained.* The former variable was categorized in the following way:

1. Teachers who wrote a Level 1 or Level 2 question as an example of a question that assesses higher-order thinking skills
2. Teachers who wrote a Level 3 or Level 4 question as an example of a question that assesses higher-order thinking skills

The latter variable was categorized in the following way:

1. Teacher who have obtained a baccalaureate degree
2. Teachers who have obtained either a master or a doctor degree

The variables for Null Hypothesis 7 were *categorization of the teacher's question* and *highest ranked educational object from Bloom's taxonomy.* The former variable was categorized in the following way:

1. Teachers who wrote a Level 1 or Level 2 question as an example of a question that assesses higher-order thinking skills
2. Teachers who wrote a Level 3 or Level 4 question as an example of a question that assesses higher-order thinking skills.

The latter variable was categorized in the following way:

1. Knowledge
2. Comprehension
3. Application
4. Analysis

5. Synthesis
6. Evaluation

Method of Data Collection

To gather the data for this study, questionnaires were mailed to 234 randomly selected full-time high school mathematics teachers in the state of Montana during the fall of 1998. As explained above, 21 randomly selected teachers were added to the sample and mailed questionnaires during the winter of 1998. The Total Design Method (TDM), as outlined by Dillman (1978), was utilized. This entailed sending

- a postcard to each teacher announcing that they had been selected for the study and would receive a questionnaire in one week;
- a cover letter and questionnaire one week after the postcard was sent;
- a postcard reminder to all 255 teachers one week after the questionnaire was sent;
- a cover letter and replacement questionnaire to all non-responders four weeks after the reminder was sent;
- a final mailing of a cover letter and replacement questionnaire to all non-responders three weeks after the second questionnaire was sent.

Non-responders from among the 21 teachers who were subsequently added to the sample received only one replacement questionnaire; not two, as were some non-responders in the original sample. Dillman examined 38 surveys that used the TDM either completely or in part and found that those surveys that used the TDM completely on a specialized population obtained an average response rate of 81%. Those that used the TDM completely on the general population obtained an average response rate of 73%. (p. 27)

The questionnaire designed for this study was targeted at the specialized population described above.

Reliability and Validity

The questionnaire for this study required teachers to provide the following information:

- Write two questions; one that they would give their students to assess higher-order thinking skills and another that they would give to assess lower-order thinking skills;
- Determine the approximate amount of professional development they have received from 1996 to 1998 and the amount of SIMMS professional development they had received from 1992 to 1998;
- List the textbooks that they use in their classroom;
- Rank-order a list of educational objectives;
- Indicate the number of years they have taught mathematics in the public schools and the total number of years that they have taught;
- Indicate the highest degree they have received and the subject area for which they received that degree;
- Indicate the number of credit hours they have accumulated from classes in assessment or evaluation at the college level.

Please refer to the copy of the original questionnaire in Appendix A.

Two aspects of reliability were addressed: First, the reliability of the questionnaire and second, the reliability of the categorization of question Q2a by the researcher. Reliability of the questionnaire was established through a test/retest conducted over a four-week period, in the winter of 1998, involving seven high school mathematics teachers. Originally, nine teachers were scheduled to participate, but two did not return the first questionnaire. The results of this test/retest are shown in Table 1.

Table 1. Results of Test/Retest

Question Number	Test	Retest	Test	Retest
	Teacher 1		Teacher 2	
Q-2a	Cooney 3	Cooney 4	Cooney 4	Cooney 2
Q-4	500 hrs.	600 hrs.	375 hrs.	320 hrs.
Q-5	SIMMS Certified	SIMMS Certified	SIMMS Certified	SIMMS Certified
Q-6	SIMMS Curriculum	SIMMS Curriculum	Mixed Curriculum	Mixed Curriculum
Q-7	Bloom 3	Bloom 3	Bloom 5	Bloom 2
Q-8	4 Years	4 Years	9 Years	9 Years
Q-9	4 Years	4 Years	8 Years	8 Years
Q-10	10,11,12	9,10,11,12	9,10,11,13	9,10,11,12
Q-11	Bachelor	Bachelor	Bachelor	Bachelor
Q-12	Math Ed.	Math Ed.	Math	Math
Q-13	4-6 Credits	7 or More Credits	0-3 Credits	0-3 Credits
	Teacher 3		Teacher 4	
Q-2a	Cooney 2	Cooney 2	Cooney 2	Cooney 2
Q-4	30 hrs.	8 hrs.	520 hrs.	520 hrs.
Q-5	SIMMS Introduced	SIMMS Introduced	SIMMS Certified	SIMMS Certified
Q-6	Mixed Curriculum	Mixed Curriculum	Non-SIMMS Curriculum	Non-SIMMS Curriculum
Q-7	Bloom 6	N/A	Bloom 6	Bloom 6
Q-8	2 Years	2 Years	7 Years	7 Years
Q-9	2 Years	2 Years	5 Years	5 Years
Q-10	9,10,11	9,10,11,12	9,10,11	9,10,11
Q-11	Bachelor	Bachelor	Bachelor	Bachelor
Q-12	Math and Other	Math and Other	Math Ed.	Math Ed.
Q-13	4-6 Credits	0-3 Credits	4-6 Credits	4-6 Credits

Table 1 Results of Test/Retest (Continued)

Question Number	Test	Retest	Test	Retest
	Teacher 5		Teacher 6	
Q-2a	Cooney 3	Cooney 3	Cooney 3	Cooney 3
Q-4	135 hrs.	348 hrs.	350 hrs.	210 hrs.
Q-5	Non-SIMMS	Non-SIMMS	SIMMS Certified	SIMMS Certified
Q-6	N/A	N/A	Non-SIMMS Curriculum	Non SIMMS Curriculum
Q-7	Bloom 6	Bloom 6	Bloom 6	Bloom 6
Q-8	18 Years	22 Years	7 Years	7 Years
Q-9	12 Years	11 Years	7 Years	7 Years
Q-10	N/A	N/A	9,10,11,12	9,10,11,12
Q-11	Master	Master	Bachelor	Bachelor
Q-12	Other	Other	Math Ed.	Math Ed.
Q-13	4-6 Credits	4-6 Credits	4-6 Credits	7 or More Credits
	Teacher 7			
Q-2a	Cooney 2	Cooney 2		
Q-4	60 hrs.	20 hrs.		
Q-5	Non-SIMMS	Non-SIMMS		
Q-6	Non-SIMMS Curriculum	Non-SIMMS Curriculum		
Q-7	Bloom 1	Bloom 2		
Q-8	4 Years	3 Years		
Q-9	4 Years	3 Years		
Q-10	10,11,12	10,11,12		
Q-11	Bachelor	Bachelor		
Q-12	Math and Math Ed.	Math Ed.		
Q-13	4-6 Credits	0-3 Credits		

There were some discrepancies between the test and retest for most of the teachers. For question Q-2a, Teachers 1 and 2 had a discrepancy. The discrepancy for Teacher 1 was one level on Cooney's taxonomy. Questions at both levels 3 and 4 assess higher-order thinking skills and therefore this discrepancy is not serious. The discrepancy for Teacher 2, on the other hand, is large (from 4 to 2). This may be explained by the fact that the teacher did not originally write a question for the retest and, when subsequently asked to, hurriedly wrote one and sent it in the mail.

There were several discrepancies for question Q-4. The differences in hours between the test and retest range from 0 to 213 hours with an average of about 81 hours per teacher, or 27 hours per year.

Teachers 2 and 7 had a discrepancy on question Q-7 while Teacher 3 failed to complete it on the retest. It was decided to rewrite this question and conduct a test/retest during the summer of 1998. This will be discussed below.

Only Teacher 5 was inconsistent on questions Q-8 and Q-9. Teacher 5 also had taught longer than any of the other teachers and may have forgotten exactly how many years she has taught.

Teachers 1 and 3 were inconsistent on question Q-10. On the first questionnaire, Teacher 1 commented that he had also taught grade 9 for three years. This may have been the reason that it was included on the retest. Teacher 3 should not have made the same mistake since he was teaching at the same time that he completed the questionnaire. It should be noted that this teacher included a comment on his retest stating that "these questionnaires are getting annoying."

Four teachers were inconsistent on question Q-13. Teachers 1 and 6 were taking courses that included an assessment component during the semester that they completed the questionnaires and may have therefore included those credits on the retest. Teacher 3 appears to have filled out the retest in a hurried manner. This may account for the inconsistency. The inconsistency of Teacher 7 cannot be explained.

After completing the questionnaire for the first time, the teachers who participated in the test/retest were asked to complete a survey evaluating the questionnaire. Teacher 1 suggested that the instructions for question Q-13 indicate that only classes in assessment or evaluation should be considered and that question Q-11 include the option "currently working on an advanced degree". This was one of the teachers who were taking courses during the semester that they completed the questionnaires. Teacher 2 thought that the option "other" should be included in question Q-5. Teacher 7 commented that the questionnaire was "a little verbose" and that this made it hard to read every word. The same teacher thought that the instructions for questions Q-4 and Q-5 should specify whether or not pre-teaching experience should be included in the number of hours of professional development. These suggestions were taken under consideration and appropriate changes were made.

As stated above, question Q-7 was rewritten and piloted during the summer of 1998. Ten students attending a graduate course for mathematics teachers (Advanced Teaching Strategies) were asked to rank-order the six items. Three weeks later they were again asked to rank-order the items. The results are shown in Table 2. Three of the ten students were inconsistent. Several of the students had trouble interpreting the meaning of the words "analyze" and "synthesize". It was therefore decide to replace these words with "breakdown" and "combine", respectively; both of which come directly from the condensed version of Bloom's taxonomy (Bloom, 1956, pp. 205-206).

Table 2. Test/Retest Results for Revised Question Q-7

	Objective Selected as First	
Student	Test	Retest
1	Bloom 6	Bloom 4
2	Bloom 5	Bloom 5
3	Bloom 6	Bloom 3
4	Bloom 3	Bloom 3
5	Bloom 3	Bloom 3
6	Bloom 3	Bloom 3
7	Bloom 3	Bloom 5
8	Bloom 3	Bloom 3
9	Bloom 6	Bloom 6
10	Bloom 5	Bloom 5

Members of the researcher's committee and mathematics educators from Montana State University determined the validity of the questions used in the questionnaire. To obtain reliability in categorizing the teacher-generated problems (question Q-2a), the researcher and his committee Co-Chair (Dr. Maurice Burke) trained in classifying exam questions using questions submitted by teachers for the test/retest described above. A rubric for categorizing questions was designed based upon Cooney's taxonomy and guidelines suggested by Bloom (1956). Dr. Burke and the researcher independently categorized 40 randomly selected teacher-generated problems from the questionnaires. After this was completed, an interclass correlation (ICC) was computed, based upon a method outlined by Futrell (1995, pp. 81-86). This method was designed to estimate the reliability of two or more judges' averaged ratings. Any value lower than a 0.7 is considered unacceptable, whereas a value higher than 0.9 is considered excellent. An ICC of 0.836 was computed for the 40 questions categorized by Dr. Burke and the researcher. Fleming and Chambers (1983) found that agreement between judges, when categorizing according to Bloom's taxonomy, could be achieved in 85% or more of the classifications through explanation and practice. The reliability of the classification of data from the other questions (Q-4 through Q-13) was obtained by designing the categories so that they were mutually exclusive.

Another form of reliability, pertaining to the classification of teacher-generated problems, must be discussed. Schrag (1989) states that "… whether a task requires higher-order thinking depends on the resources of the thinker." (p. 530) Before we can categorize a problem using either Cooney's or Bloom's taxonomy, we must know what the students already knows and what types of problems were assigned to the students by the teacher. The problem was addressed in this research by requiring the teacher to indicate the specific class and curriculum unit for which the question was designed. Both Dr. Burke and the researcher found this information vital for categorizing the problems and helpful in resolving most differences in categorization encountered during the test of reliability.

The validity of Cooney's taxonomy follows from the validity of Bloom's taxonomy. Cooney's taxonomy has the following categories:

<u>Level 1.</u> Simple computation or recognition;

<u>Level 2.</u> Comprehension. Students must make some decision but once made the solution process is straightforward, e.g., a simple one step problem;

<u>Level 3.</u> An application or multistep problem in which [the] student must make several decisions about [the] solution process or what operations to use; and

<u>Level 4.</u> Nonroutine or open-ended problem. (Cooney, 1992, p. 7)

This taxonomy was developed directly from Bloom's taxonomy (Bloom, 1956), but focuses only on mathematical objectives. Level 1 is associated with Bloom's category *Knowledge*, Level 2 is associated with Bloom's category *Comprehension*, Level 3 is associated with Bloom's category *Application*, and Level 4 encompasses Bloom's three categories: *Analysis*, *Synthesis*, and *Evaluation*. Hence, the validity of this taxonomy is directly related to the validity of Bloom's taxonomy. Furthermore, this taxonomy has already been applied to research on teacher-made questions and the results were published over five years ago.

As noted above, the researcher designed a rubric to help categorize teachers' questions using Cooney's taxonomy. This rubric (see Appendix B) was designed using Bloom's taxonomy (1956) as a guide and was expanded and refined based upon the categorization of teacher-generated questions by both Dr. Burke and the researcher. This process was completed before the reliability test, discussed above, was administered.

Research Design and Analytical Techniques

There are both rural and non-rural schools in Montana with varying percentages of Native American students. These variables were controlled through the use of a stratified random sample.

Each null hypothesis in this study was tested with an $\alpha = .05$. This, rather than an $\alpha = .01$, was selected since the SIMMS IM curriculum and SIMMS professional development programs have already been introduced into the state of Montana and the costs for their design and development have already been accrued. Therefore, the decision by any school system to incorporate the SIMMS IM curriculum as a result of this study will be no more expensive than the decision to incorporate any other new curriculum.

The following null hypotheses were tested using the chi-square test of independence:

3. There is no relationship between the categorization of teachers' written examples of questions that test for higher-order thinking skills and the amount of SIMMS professional development that they have received.
4. There is no relationship between the categorization of teachers' written examples of questions that test for higher-order thinking skills and the use of the SIMMS IM curriculum in their classroom.
5. There is no relationship between the categorization of teachers' written examples of questions that test for higher-order thinking skills and the number of credit hours that they have accumulated from college classes in assessment or evaluation.
6. There is no relationship between the categorization of teachers' written examples of questions that test for higher-order thinking skills and the highest academic degree they have obtained.
7. There is no relationship between the categorization of teachers' written examples of questions that test for higher-order thinking skills and which educational object from Bloom's taxonomy they rank most important in their classroom.

The following null hypotheses were tested using an analysis of variance (ANOVA):

1. There is no significant difference between the mean number of hours of professional development for teachers who wrote Level I, Level II, Level III, or Level IV questions as examples of questions that test for higher-order thinking skills.
2. There is no significant difference between the mean number of years teaching mathematics for teachers who wrote Level I, Level II, Level III, or Level IV questions as examples of questions that test for higher-order thinking skills.

Qualitative Research Methodology

Purpose of the Qualitative Research

The purpose of the qualitative research was to assist in the interpretation and explanation of the quantitative data. Generating hypotheses as to why teachers assess the cognitive levels that they do in the classroom helped the researcher interpret the statistical results obtained from the data provided by the questionnaire.

Population and Sampling Procedure

The population for the qualitative research was the 59 teachers who indicated on their return postcard that they would be willing to participate in an interview. A stratified sample (the same stratification as used for the quantitative portion of this study) of those 59 teachers was used for this phase of the research. Only two teachers from schools with a substantial Native American population would agree to be interviewed.

Methods of Data Collection

Data was collected through telephone interviews and by obtaining at least three tests from each teacher that they had actually used in their classroom. The interview protocol was piloted on two teachers from Belgrade High School and one teacher from Park Senior High School in Livingston. Below is a list of the questions that each teacher was asked to respond to during the interview.

1. Describe the types of test questions used by your high school mathematics teachers to assess your understanding of mathematics.
2. Describe the types of test questions used by your college mathematics teachers to assess your understanding of mathematics.
3. List one, two, or at most three factors that have most influenced the way that you assess your students and explain how they have influenced your assessment practices.
4. To gain more detail about your assessment practices in the classroom, describe to me other methods than regular classroom tests that you use to assess your students understanding of mathematics.
5. In order of importance, list which methods of assessment provide the most weight when you eventually assign a letter grade to a student's performance in the classroom.
6. Presently, there is discussion in the education journals of assessment of higher-order thinking skills, relative to Bloom's taxonomy. As an expert in