THE EFFECT OF ITEM FORMAT ON
MATHEMATICS ACHIEVEMENT TEST SCORES

by

Leslie Hubert Dukowski

B. Sc., University of British Columbia, 1973

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS

in the Faculty
of

Education

We accept this thesis as conforming to the
required standard

THE UNIVERSITY OF BRITISH COLUMBIA

April, 1982

Department of Mathematics Education

The University of British Columbia
1956 Main Mall
Vancouver, Canada
V6T 1Y3

Date 82.04.06

Abstract

Thesis Supervisor: Dr. David F. Robitaille

The purpose of the study was to determine whether item format significantly affected scores on a mathematics achievement test. A forty-two item test was constructed and cast in both multiple-choice and constructed-response formats. The items were chosen in such a way that in each of three content domains, Computation, Application, and Algebra, there were seven items at each of two difficulty levels. The two tests were then administered on separate occasions to a sample of 213 Grade 7 students from a suburban/ rural community in British Columbia, Canada.

The data gathered was analysed according to a repeated measures analysis of variance procedure using item format and item difficulty as trial factors and using student ability and gender as grouping factors. Item format did have a significant ($p < 0.05$) effect on test score. In all domains multiple-choice scores were higher than constructed-response scores. The multiple-choice scores were also transformed using the traditional correction for guessing procedure and analysed. Multiple-choice scores were still significantly higher in two of the three domains, Application and Algebra. There were significant omnibus *F*-statistics obtained for a number of interactions for both corrected and uncorrected data but there were significant Tetrad differences ($p < 0.10$) only for interactions involving format and difficulty.

The results indicate that students score higher on a multiple-choice form of a mathematics achievement test than on a constructed-response form, and therefore the two scores cannot be considered equal or interchangeable. However, because of the lack of interactions involving format, the two scores may be considered equivalent in the sense that they rank students in the same manner and that the intervals between scores may be interpretable in the same manner under both formats. Therefore, although the traditional correction for chance formula is not sufficient to remove differences between multiple-choice and constructed-response scores, it may be possible to derive an empirical scoring formula which would equate the two types of scores on a particular test.

CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

ACKNOWLEDGEMENT

I wish to thank the members of my thesis committee, Dr. David Robitaille, Dr. Tom O'Shea, and Dr. Todd Rogers for their guidance. It has been a privilege for me to work with them.

In addition I would like to express my gratitude to the Grade 7 students and teachers in Langley School District who participated in the study. I am pleased to have been associated with colleagues of their professional stature.

Finally, I would like to thank my family for their encouragement, especially my wife Lesley for her patience and constant support.

Chapter 1

BACKGROUND

During the past decade, educators have witnessed a movement toward minimum-competency testing and large-scale evaluation of educational programs. In British Columbia, assessments in Reading, Writing, Mathematics, Science, Physical Education, and Social Studies have been conducted. Similar to comparable state and provincial assessments, the goal of the B.C. Learning Assessment Program is not to measure individual student performance, but to provide information about student learning on a province-wide basis. That is, the intent is to measure the extent to which the basic objectives of the educational system are being achieved by all students. On the other hand, the intent of the former provincial examination system appeared to be to determine the percentage of students which should be admitted to higher schooling (Mussio and Greer, 1980, pp. 26-27).

Once data have been collected and analysed, some judgment must be made regarding the acceptablity of students' performance. A major component of the B.C. Learning Assessment Program is the interpretation of the results obtained by students on the assessment tests. This interpretation process is not without difficulties. Mussio and Greer (1980, p. 35) have discussed the concern over the confusion of norms and standards when interpreting assessment data.

For example, Mathematics and Science Assessments utilized items from the National Assessment of Educational Progress in the United States and B.C. students outperformed U.S. students on a number of these items. On the basis of such evidence from a norm-referenced standpoint, one might conclude that the students in the schools are achieving the goals of the curriculum at a satisfactory level. From a criterion-referenced standpoint, however, the decision is not at all clear. Even though the students outperformed their American counterparts, it still may be that the level of achievement on certain basic skill items was unacceptably low. Therefore, because the Assessment program is intended to determine whether or not students have achieved the objectives of the curriculum, a traditional, norm-referenced approach is not suitable. Mussio and Greer (1980) point out, however, that

> Experience dealing with interpretation panels, involving both educators and members of the public, for six assessments has repeatedly demonstrated an initial skepticism [on the part of the interpretation panels] of any method of interpretation that is other than normative. (p. 35)

It is not surprising that people are skeptical of criterion-referenced interpretation procedures. In the past, educators, and particularly those involved with educational measurement, have tended to stress procedures which measure relative rather than absolute worth (Burton, 1972, p. 1).

As a part of the 1981 Mathematics Assessment, criterion-referenced interpretations were made of student achievement on multiple-choice items (Robitaille, 1981). Mason (1979) claims that if provincial assessments are intended to measure essential, or core curriculum, learning objectives and if these objectives reflect real-life skills, then multiple-choice items are not appropriate. He points out that there are very few real-life situations where one is required to select a response from a variety of options. More often, one is required to construct a response.

Mason also claims that the type of thinking required to answer multiple-choice questions is different from that required to answer constructed-response items. For example, on a multiple-choice item it may be possible for respondents to choose the correct answer by guessing, eliminating alternatives judged unreasonable, or working backward from the given answers. These strategies are of little use on constructed-response questions. The scores from multiple-choice and open-ended forms of an achievement test may rank students in essentially the same manner, however the scores may not be interchangeable when one wishes to make criterion-referenced judgments (Mason, 1979, p. 11).

Although these concerns regarding the use of multiple-choice items appear sound, Bracht and Hopkins' (1968) review of the literature led them to conclude that many of the differences of opinion regarding the content validity of objective tests, including multiple-choice tests, are not based on empirical evidence. With regard to the question of whether or not the cognitive processes involved in recall and recognition are the same, Tulving and Watkins (1973) claimed that the same psychological processes underlie both activities. Traub and Fisher (1977) also claimed that both constructed-response and multiple-choice forms of tests of mathematical reasoning measure the same attributes.

Statement of the Problem

The 1981 British Columbia Mathematics Assessment sought to determine the attainment of curriculum objectives by the entire student population at Grades 4, 8, and 12 (Robitaille, 1981). The technical considerations of large-scale testing made the use of multiple-choice items preferable to constructed-response items. Following the administration and scoring of the Assessment instruments, an attempt was made to arrive at criterion-referenced judgments based on the test item results. There have been concerns expressed that the scores obtained by students on multiple-choice items may not be interchangeable with, nor equivalent to scores obtained on the same items in constructed-response format. This being the case, the judgments made by the Interpretation Panels may be suspect. This study sought to investigate whether or not there were differences in the character of the scores which would seriously affect the interpretation of item results.

A demonstration that the two scores measure the same attributes in essentially the same way may relieve some of the apprehension on the part of those interpreting Assessment data. If the scores are shown not to be equivalent measures, then a description of the precise nature of the differences between the scores may make the interpretation more meaningful.

The general questions to be addressed by the study, then, arise from the concerns as to whether or not multiple-choice test item results are interpretable in the same way as constructed-response item results. The specific questions are as follows:

1. Do students score higher on a multiple-choice Mathematics achievement test than on the same test in constructed-response format?

2. Do boys outperform girls on Mathematics achievement tests in either format?

3. Is there an interaction between format and gender, and if so, what is the nature of the interaction?

4. Is there an interaction between format and ability, and if so, what is the nature of the interaction?

5. Is there an interaction between format and item difficulty, and if so, what is the nature of the interaction?

6. Are there any more complex interactions which may affect test score, and if so, what are the natures of those interactions?

Definition of Terms

The following terms are used throughout the study and are defined here for convenience.

Recognition items are those items for which the respondent chooses an alternative from a given list of choices. Recognition items will also be referred to as multiple-choice items.

Recall items are those items for which the respondent must construct a response. Recall items will also be referred to as constructed-response items.

Objective tests are those which contain recognition items or those recall items which require only single word or single phrase responses. That is, there is a clearly defined right answer and there is a dichotomous decision made on the part of the grader as to whether or not the response is acceptable.

Essay-type tests are those which contain items to which the respondent must answer using more than one phrase or sentence. Although there may be well defined criteria for the acceptability of responses, the grader rarely makes a dichotomous decision as to whether a response is acceptable.

Response sets have been defined in the following way:

> A response set is defined as *any tendency causing a person to give different responses to test items than he would when the same content is presented in a different form.* (Cronbach, 1946, p. 476, italics in the original)

For example, a person may have a tendency to agree with statements framed in a positive way and to disagree with statements framed in a negative way. With such a response set, a person might agree with content in one context, framed positively, and disagree with the same content in a different context, framed negatively.

Formula scoring is the procedure by which student scores on tests containing multiple-choice items are adjusted to correct for guessing on the part of respondents.

The usual formula for correcting individual scores is:

$$S = R - (W/(k - 1))$$

where R is the number of correct responses,

W is the number of incorrect responses,

k is the number of answer options on a single item, and

S is the score corrected for guessing.

This formula is used under the assumption of random guessing on the part of respondents who do not know the correct answer.

Content domain A content domain is a body of material defined by a set of learning outcomes. The three content domains into which test items were grouped in this study were Computation, Application, and Algebra. These content domains are operationally defined in Chapter 3.

Organization of the Following Chapters

A review of the literature pertaining to the experimental questions, a description of methodology, the results of statistical analysis and a discussion of the findings of the study are found in the following Chapters. The research hypotheses are presented at the end of Chapter 2 following the review of literature. Chapter 3 contains the details of sample selection, development and administration of test instruments, and methods of data analysis. The results of the descriptive and inferential analyses are discussed in Chapter 4 and summarized in Chapter 5 which also contains a discussion of the findings of the study and their implications. The test instruments and directions for test administration are included in the appendices following the references.

Chapter 2

REVIEW OF RELATED LITERATURE

The review of the literature is organized in four sections. First, literature pertaining to the issue of recall versus recognition is presented. Next, the literature regarding response sets is discussed. Literature related to formula scoring and corrections for guessing is then reviewed and followed, finally, by a review of studies of sex-related differences in Mathematics achievement.

Recall and Recognition

One of the major questions in educational testing is that of whether or not recall and recognition tests are equally effective in measuring student ability. Proponents of the use of multiple-choice items cite the technical advantages of recognition items whereas advocates of constructed-response items point out that recall items may be used to assess partial knowledge (Cronbach, 1970, pp. 30-32).

Advantages and Disadvantages of Recall and Recognition Items

Stanley and Hopkins (1972, p. 236) claim that the "multiple-choice form is usually regarded as the most valuable and most generally applicable test form." Cronbach (1970) points out that one of the criticisms of multiple-choice tests is that they tend to be restricted to low-level thinking. He then goes on to claim that it is possible to construct multiple-choice tests which require a great deal of understanding and the application of higher cognitive processes. This opinion that multiple-choice items can test both simple and complex learning outcomes is shared by other measurement specialists as well (Gronlund, 1968, p. 26; Ebel, 1979, pp. 56-57).

The advantages of constructed-response or recall items are listed by Stanley and Hopkins (1972) as the following:

1. They are familiar to most children as they are commonly used on teacher made tests.

2. They almost completely eliminate guessing.

3. They are particularly suited to arithmetic and the physical sciences where computations are required.

They claim that a disadvantage of items which require single word or short phrase responses is that they tend to measure only factual knowledge. Gronlund (1968, p. 26), however, claims that essay questions are often used when more complex learning outcomes requiring unique responses are assessed. Ebel (1979, pp. 56-57) states that it is a misconception that item type indicates the ability tested. Rather, good essay and objective tests can require the same kind and level of ability.

According to Ebel (1979, p. 57), multiple-choice tests are more difficult to construct than essay tests. However, they can be more rapidly and reliably scored than essay tests. Stanley and Hopkins (1972, p. 218) also claim that even single-word constructed-response items are time consuming to score and not always entirely objective.

Reliability and Validity Studies

Most of the empirical studies of objective and essay tests have been concerned with reliability (Bracht and Hopkins, 1968, p. 3). For example, Kinney and Eurich (1932) tabulated the results of thirteen studies comparing recall, multiple-choice, true-false, and essay-type examinations. In six of the nine studies comparing the reliabilities of multiple-choice and constructed-response tests, the constructed-response tests were shown to have higher reliabilities; however no tests were performed to determine whether the reliabilities were significantly different. In contrast more recent results support the claim that objective tests consistently show higher reliabilities than essay tests (Bracht and Hopkins, 1968).

With regard to the validity of objective and essay tests, Bracht and Hopkins (1968) reviewed a wide range of studies designed to compare content validity. Although there was considerable difference of opinion among the authors of the studies reviewed, Bracht and Hopkins concluded that the evidence supported the contention that both types of tests measure the same things. With reference to Mathematics, Cronbach (1970, p. 31) cited College Entrance Examination Board data which showed that the results from multiple-choice and constructed-response questions had essentially the same correlations with grades in later courses in Mathematics.

Mental Processes Involved in Recall and Recognition

Mason (1979), among others, has expressed the opinion that responses to recall and recognition items do not require the same mental processes. Among psychologists, there appear to be two competing theories to describe the processes involved in recall and recognition (Rabinowitz, Mandler, & Patterson, 1977). The two theories are the unitary strength theory and the generation-recognition hypothesis. The unitary strength theory asserts that both recall and recognition have accompanying associative stimulus-response strengths. According to the theory, recognition items generally have more strength to elicit a correct response whereas recall items may not have as much stimulus information, that is, enough strength, to cause a respondent to produce a correct response.

On the other hand, the generation-recognition hypothesis asserts that while recognition requires a simple decision process, recall requires the generation of possible responses and then the elimination of unsatisfactory candidates. The decisions made in the elimination of candidates are essentially the same as those made in recognition.

Rabinowitz, et al. (1977) reported the results of a number of studies designed to provide information regarding recall and recognition of information, specifically, lists of words. These studies provided data on recall tests preceding or following recognition tests, the effect of specific instructions in recalling words from the lists provided, the strength of items, and the effects of using a taxonomy when recalling or recognizing words. Based on an analysis of the results, the authors claimed that recall and recognition made use of similar mental processes.

Similarly, Tulving and Watkins (1973) claimed that the same psychological processes underlie both recall and recognition. They reported a study in which subjects were shown a number of sequences of five-letter words. Following each sequence, the subjects were asked to reproduce the sequence. Cues of one, two, three, four, and five letters were given on each test except one, where no cues were provided. Performance improved in direct relation to the number of cues given. However, there were no discontinuous jumps in performance even though, as the authors claimed, for example, the task of generating probable alternative words from three-letter cues is much more difficult than from four-letter cues. This continuity of performance in relation to the number of cues provided led the authors to believe that there is no clear distinction between recall and recognition. Moreover, they felt that the hypothesis that the two are continuous is a more useful construct than the proposition that the two are disjoint processes.

Bracht and Hopkins (1970) reported a study using sophomores enrolled in an educational psychology course. During the semester, students were administered both essay and multiple-choice examinations. Each examination was designed to measure higher cognitive processes and great care was taken to control for confounding factors such as answer length and penmanship when grading the essay questions. Following an analysis of the data, the authors reported that the assumption that multiple-choice and essay tests measure different variables was not supported.

Traub and Fisher (1977) conducted a study in which they investigated whether or not tests which differed only in response format measured the same attributes. Two sets of mathematical reasoning tests and two sets of verbal comprehension tests were administered to Grade 8 students under three different formats. The tests were presented in constructed-response, multiple-choice, and Coombs multiple-choice format. In the Coombs format, respondents were required to identify incorrect options. Based on the data obtained, the authors claimed that the tests of mathematical reasoning measured the same attributes regardless of format. The tests of verbal reasoning were found not to be equivalent.

Burton (1972) conducted a study to investigate the effects of item-scoring formulas. In her study items which differed only in response format, multiple-choice or constructed-response, were administered. The results led her to speculate that there was an inherent difference between the two types of items. In particular, differences in student achievement ranged between five and fifteen percent in in favour of the multiple-choice items (pp. 129-130). However, she went on to state that for items which contained very plausible distractors, this difference did not appear, and in most cases the interpretation of the data was not affected by the supposed inherent difference.

Response Sets

Response sets cause subjects of equal ability to consistently give differing responses according to irrelevant characteristics of test items. Consequently, response sets are a factor which must be considered when constructing achievement tests.

Characteristics of Response Sets

Cronbach (1946) lists the following response sets:

1. Guessing.

2. Acquiescence: some people, when faced with a choice about which they are not sure, tend to answer positively; others tend to answer negatively.

3. Speed versus accuracy: students who work quickly, guessing on items, may receive undeserved higher scores in comparison to their slower, more thoughtful classmates.

4. Definition of judgment categories: the categories used on a test, such as like or dislike, may mean different things to different people.

5. Inclusiveness: some people may be more inclined to include many answers or points in their responses while others may be more selective or limited in the points made in an essay or choose a smaller number of alternatives on multiple-choice tests.

6. Response sets on essay tests: some of these sets are inclusiveness, style of composition, and degree of organization.

Stanley and Hopkins (1972) list the first three categories above and also add a category, positional preference: the tendency of some respondents to consistently choose the option appearing in a particular place in the list of possible answers. Stanley and Hopkins go on to point out, however, that recent research has failed to confirm this response set.

Having categorized response sets, Cronbach (1946), lists these characteristics of response sets:

1. They are reliable from test to test.

2. They increase test reliability.

3. They raise or lower test validity depending on whether or not the response sets are correlated with the criterion.

4. They always lower content validity.

5. They have the most effect on difficult items.

6. They have the greatest influence in situations perceived by the respondent as ambiguous or unstructured.

7. They appear to be uncorrelated across subject fields. That is, a respondent may gamble on Mathematics tests but not English tests, or vice versa.

8. They interfere with inferences made on the basis of the results of tests.

A study reported by Hopkins (1964) contains a literature review which supports the claims about the characteristics of response sets given above. In addition, Hopkins also claimed, on the basis of the review, that response sets have been found to be relatively independent of ability, but related to personality.

Sherriffs and Boomer (1954) conducted a study which examined the relationship of guessing behaviour to personality. In the study, the subjects were told that the scores on a true-false examination would be computed by subtracting the number of wrong responses from the number of correct responses. The results of the study showed that introverted subjects with low self-esteem and high concern with the impression that they make on others tended to be penalized by a correction for guessing when their scores were compared to those of other students. In other words, these students were less likely to gamble and guess answers.

Effects of Response Sets and Guessing

It has been stated above that response sets have an effect on test reliability and validity. In a study conducted by Hopkins (1964), grade equivalent scores on standardized multiple-choice mathematics achievement tests were investigated and the scores on equivalent forms of the tests in constructed-response format were compared with the multiple-choice scores. The results showed that by answering all items and attaining chance success, that score could be interpreted as an acceptable grade equivalent score. Hopkins went on to explain that the reliability of such standardized tests may be due to the speed versus accuracy set. That is, although the instruments may appear to test content objectives, and yield high reliability coefficients, they may indeed be measuring irrelevant but stable factors.

In an assessment based on multiple-choice items, then, those interpreting the results will be faced with the problem of scores possibly inflated by guessing. Thorndike (1971) defines guessing as a "loose, general term for an array of behaviors that occur when an examinee responds to an alternate choice question to which he does not 'know' the answer" (pp. 59-61). Thorndike then goes on to list some behaviours that occur during guessing:

1. judging some answer choices to be wrong and selecting a response from the remaining alternatives;

2. using unintended semantic and syntactic cues in the wording of the responses or the question stem;

3. being misled by plausible but wrong responses constructed by the item writer;

4. making an unsure response based on an attractive element in one of the choices;

5. responding in a random fashion, using a pattern of responses or some specific response pattern.

Rowley and Traub (1977) conducted a short review of formula scoring literature. Their review focussed on the assumptions regarding pupil behaviour during examinations. The evidence suggested that the tendency to guess is correlated with personality and therefore introduces a confounding effect. However, "Do not guess" instructions may introduce a confounding influence as well because there may be differential compliance with the instructions according to personality characteristics.

Rowley and Traub (1977) also claimed that students tend to score higher than chance by "just guessing". This assertion is only weakly supported by the results of a study conducted by Ebel (1968). In that study students responded to true-false items and indicated which of the responses were simply blind guesses. Only a small proportion of the responses (3%-8%) were reported as blind guesses and those blind guesses were between 52% and 56% correct, only slightly better than the proportion of scores expected to be correct by chance alone. Ebel (1979, p. 201), however, encouraged students to make rational guesses on objective items.

Burton's (1972) review of the literature on guessing showed that some changes in format, notably the inclusion of the "I don't know" alternative, may reduce guessing. She was of the opinion, though, that there is no way to prevent all guessing on multiple-choice achievement tests.

Formula Scoring and Corrections for Guessing

A great deal has been written about formula scoring and the effects of correction formulas on test reliability and validity. Rowley and Traub (1977) reported that the results of empirical studies are equivocal and that, on unspeeded tests, the decision to use or not use formula scoring is a value judgment.

Ebel (1979, pp. 194-8) claimed that both corrected and uncorrected scores rank students in essentially the same order and the probability of achieving a respectable score on a good objective test by guessing is slight. Also, if examinees are well-motivated and have time to attempt all items, the effects due to guessing will be reduced. In addition, it is not poor practice to encourage students to make rational guesses, the results of which may provide information on the general achievement of the students. Finally, a guessing correction may remove the incentive for slower students to guess on speeded tests but the corrected scores may be contaminated by the action of response sets.

Lord and Novick (1968, pp. 302 ff) described more complicated scoring formulas than the most commonly used S = R - (W/(k - 1)). Some of these formulas have been developed in an attempt to evaluate partial knowledge and/or make use of weighting the scores assigned to certain items or response choices within the item to minimize mean square error. These scoring weights are determined empirically from the test data. Burton (1972) compared a number of scoring methods. One of the methods was simple number right and another was the simple correction given above. The other eighteen methods investigated employed two formulas which used empirical data to assign scoring weights and 16 different combinations of formulas based on a complicated series of decision procedures.

Lord and Novick (1968) suggest that researchers and theoreticians are unlikely to abandon the search for more refined methods to glean an increased amount of information from test scores. They point out however that

> . . . what little experimental work has been done in the traditional methods of formula scoring has not been encouraging, and that no experimental work has been published that supports the new methods. Thus, at present, the sole recommendation of these new methods is their strong conceptual attractiveness. In evaluating any new response method, it will be necessary to show that it adds more relevant ability variation to the system than error variation, and that any such relative increase in information retrieved is worth the effort . . . (p. 314)

Educators and measurement specialists are divided on the issue of formula scoring. Lord (1975) stated that "Religion, politics, and formula scoring are areas where two informed people often hold opposing views with great assurance" (p. 7).

The central assumption made when employing the most common corrections for guessing is that respondents will guess randomly when faced with an item to which they do not know the answer with absolute confidence. Lord (1975) showed that under this assumption, both formula scoring and simple number right give unbiased estimates of the same quantity. However, Lord (1963, 1975) also suggested that the assumption of random guessing is indefensible.

Sex-related Differences in Mathematics Achievement

Swafford (1980) reported that while the literature of the 1960's and 1970's generally held that sex-related differences in mathematics achievement did not appear until adolescence, more recent studies have shown that these differences are negligible at all ages when the number of years of mathematics studied by the subject is controlled. Similarly Wolleat, Pedro, Becker, and Fennema (1980) claimed that while research on cognitive factors has been inconclusive, studies examining non-cognitive factors have yielded interesting results. In particular, females have been found to be less confident than males about their ability in mathematics and tend to underestimate their ability. Compared with males' beliefs, females believe that mathematics will be less useful to them in the future. These factors seem to have caused females to avoid senior courses in mathematics.

The results of differences in achievement between males and females have been discussed in the General Reports from both of the British Columbia Assessments of Learning in Mathematics (Robitaille and Sherrill, 1977; and Robitaille, 1981). The results of the first B.C. Assessment and of the 1979 NAEP Mathematics study are also described by Erickson, Erickson, and Haggerty (1980). The three sets of assessment data show some common trends. In the primary years, males tend to outperform females on measurement items and females tend to outperform males in computation. In junior high school and beyond, males outperform females in all areas except computation.

The concern over sex-related differences is evidenced by the number of programs and projects developed in response to the demonstrated sex-related differences. Erickson et al. (1980) and Fennema, Wolleat, Pedro, and Becker (1981) have described some of these programs and commented on their effectiveness. In addition, the recommendations made as a result of the 1977 and 1981 B.C. Mathematics Assessments have included some relating to sex-related differences (Robitaille and Sherrill, 1977; Robitaille, 1981).

Hypotheses

The review of literature gives rise to the following hypotheses.

Given two Mathematics achievement tests containing the same items, one using multiple-choice format and the other using constructed-response format,

1. There is no difference in the mean score obtained by Grade 7 students on the two forms in each of the content domains considered.

2. There is no difference between the test scores obtained by Grade 7 males and Grade 7 females on either form of the achievement test for each of the content domains considered.

3. There is no difference between the test score obtained by the Grade 7 students of differing abilities on either form on the achievement test in each of the content domains considered.

4. There are no interactions involving item format and item difficulty, item format and gender, or item format and student ability which have significant Tetrad differences.

Chapter 3

METHOD

The purpose of this study was to investigate the effect of item format on achievement test score. In order to measure this effect a group of Grade 7 students was administered a Mathematics achievement test on two occasions. On the first occasion half the students completed a test consisting of 42 items in multiple-choice format, while the other half completed the same test with the items in constructed-response format. Two weeks later the students were administered the tests once again. On this occasion those who had previously written the multiple-choice test responded to the test in constructed-response format, and those who had written the constructed-response test on the first occasion, wrote the test in multiple-choice format. The tests were subsequently scored, and the test scores analysed.

A description of the test development, sample selection, and pilot testing are found below. The details of test administration, test scoring, and data analysis are also presented.

Development of the Tests

The following section contains a description of the origin of the test items and the content domains. The procedures used to pilot the tests are also described.

Origin of the Test Items

During the summer of 1981, three graduate students in Mathematics Education at U.B.C., including the writer, were contracted to construct multiple-choice achievement test items for the 1981 B.C. Mathematics Assessment (Klassen, Dukowski, and deGroot, 1981). Test objectives were determined for each of the three grades (4, 8, and 12) involved in the Assessment, and pools of items were constructed for each objective for each grade. During the course of development, the items were reviewed by the Contract Team for the Assessment and also by the Assessment Advisory Committee. This Advisory Committee consisted of educators from the schools and colleges, Ministry of Education personnel, members of B.C. Research (the technical agency for the Assessment program), and a school trustee (see Robitaille, 1981 for further discussion). Pilot tests were constructed from the pools of items and administered to Grades 4, 8, and 12 classes in November, 1980. The items used for this study were selected from those piloted items for Grade 8.

Content Domains

The items on the tests used in this study were grouped into three content domains, Computation, Application, and Algebra. These content domains were defined so as to include only core material, that is, essential learning for all students, from the British Columbia mathematics curriculum for Grades 7 and 8. The prescribed content for Grades 7 and 8 Mathematics in British Columbia schools is described in *Mathematics, Curriculum Guide Years One to Twelve* (B.C. Ministry of Education, 1978, pp. 22-26). In the *Guide* the learning outcomes for Grades 7 and 8 are grouped under eight strands.

I. Set and set operations

II. Number and number operations

III. Geometry

IV. Measurement

V. Problem Solving

VI. Graphs and functions

VII. Applications of mathematics

VIII. Logical thinking

For this study the writer chose objectives from the strands and grouped them in three content domains. As mentioned above, only those objectives in the Guide designated as essential learning for all students were considered for inclusion in a content domain. The objectives in each content domain and their relationship to the strands in the Guide are described below.

Computation

The Computation content domain is defined by the following learning outcomes.

The student is able to

1. add, subtract, multiply, and divide whole numbers, common fractions, and decimal fractions
2. compare fractions
3. convert among common fractions, decimal fractions, and percent
4. calculate with percent.

All of the material in the Computation domain is classified in the Number and Number Operations strand of the Curriculum Guide.

Application

The Application content domain is defined by the following learning outcomes.

The student is able to

1. apply computational skills to solve word, or story problems

2. use skills with percent, ratio, and proportion to solve word, or story problems.

These objectives are categorized under VI. Problem Solving and VII. Applications of Mathematics in the Curriculum Guide.

Algebra

The Algebra content domain is defined by the following learning outcomes.

The student is able to

1. solve simple open sentences

2. translate verbal statements into expressions or open sentences

3. evaluate expressions.

These objectives are categorized under II. Number and Number Operations and V. Problem Solving in the Curriculum Guide.

Selection of Test Items

There were four criteria which governed the selection of test items:

1. The items had to be such that they could be stated in both multiple-choice and constructed-response formats with the same item stem.

2. The items had to test content in one of the three content domains--Computation, Application, or Algebra--as defined for in the study.

3. There had to be an equal number of items for each difficulty level considered. The two levels of difficulty were high difficulty ($0.375 < p < 0.500$) and low difficulty ($0.625 < p < 0.750$). Items were classified based upon difficulty levels obtained from pilot testing for the 1981 B.C. Mathematics Assessment.

4. The number of items testing content in each domain had to be equal yet the total number of items had to be such that the total test administration time would not exceed one hour.

Using these criteria, 42 items were selected from the items pilot tested in November 1980 for the 1981 B.C. Mathematics Assessment. There were seven items chosen for each of six subtests: Computation High Difficulty, Computation Low Difficulty, Application High Difficulty, Application Low Difficulty, Algebra High Difficulty, and Algebra Low Difficulty. For each subtest, two test forms were constructed: multiple-choice and constructed-response.

The items were then randomly distributed throughout the test with the restriction that the first two items were Computation Low Difficulty items. The two forms of the test were identical except that whereas on the multiple-choice form the students selected one of five answer options, including "I don't know", following each item stem; on the constructed-response form of the test, the same item stems, in the same order, were followed by a line upon which the students recorded their answers. Students responded directly in the test booklets which are reproduced in Appendix A.

In order to verify the content validity of the test items, two of the investigator's colleagues, both experienced mathematics teachers, were given descriptions of the content domains and they independently classified the items according to domain. There was unanimous agreement as to item classification.

Pilot Testing

The test forms were piloted in March, 1981 in the investigator's Grade 8 Mathematics class. All but one of the 27 students completed the test within 50 minutes. No problems were encountered during the administration of the tests.

Item analysis of this pilot test data conducted using the computer program LERTAP 2.0 (Nelson, 1974) revealed that the items were considerably easier than would have been expected on the basis of the Assessment pilot data. For example, the Assessment pilot data indicated that the mean difficulty for items on the Computation Low Difficulty multiple-choice subtest should be approximately 0.68; the mean difficulty obtained on the study pilot was 0.93. The tests were then piloted in the two Grade 7 classes of a neighbouring elementary school. As in Grade 8, the administration time was less than one hour. An analysis of the test results showed the item difficulties at Grade 7 to be closer to those obtained in the Assessment pilot. (The item difficulties are summarized in Table 4.2 found in the next chapter.)

The reason for the discrepancy between the item difficulties may be due to the fact that the Assessment pilot was conducted in November, while the pilot for this study was performed in March. It seems reasonable to expect that the skills of Grade 8 students would have improved over the intervening four months and that they would do better on the test items. Because the item difficulties computed from the Grade 7 pilot data were closer than the Grade 8 pilot data to those required, the study was performed using Grade 7 students.

Sample Selection

Description of the Population

The sample used in the study was selected from the population of intact Grade 7 classes of School District 35, Langley, British Columbia. Langley is a suburban-rural community located approximately 50 km from Vancouver, B.C. In contrast to more well-established school districts in the Lower Mainland, Langley is experiencing growth in student population. A wide range of socio-economic levels is represented in the community. Many Langley residents commute to blue-collar and white-collar jobs in Vancouver, and there is a sizeable number of families for whom farming is a primary or secondary source of income. The Grade 7 population consisted of 1017 students in 28 elementary schools, 20 of which had full Grade 7 classes enrolled. The other eight schools had only split Grade 6/7 classes.

Selection Technique

Of the 20 schools which had full classes of Grade 7 enrolled only the 18 schools which had a population of at least 25 Grade 7 students, excluding the school in which the tests had been piloted, were considered for participation in the study. The director of elementary instruction for the school district provided a list of those six schools which he felt were representative of the population and which had principals who were likely to agree to participate in the study. The six principals were contacted by telephone. Only one of the six principals declined to participate.

The five schools which took part in the study enrolled a total of 237 Grade 7 students in nine classes. Of these 237 students, 24 children failed to write one or both forms of the test due to absence on one or both of the testing dates. None of the data from these students were included in any of the analyses.

Subjects were grouped into three ability levels according to IQ scores as measured by the Canadian Cognitive Abilities Quantitative Battery (Thorndike, Hagen, and Wright, 1974). This Battery was administered to Grade 7 students in Langley in the Fall of 1980. Of the 213 students who wrote both tests, IQ scores were available for 191 of them. These scores were used to partition the sample into low, average, and high ability groups of roughly equal size.

Test Administration

The repeated measures design of the study required that each student respond to both forms of the test. For this reason, two testing periods were required. In order to minimize memory effects, a two-week interval separated the two testing periods. A two-week period was also used by Traub and Fisher (1977) to separate testing periods in a similar study.

The tests were first administered to students during the week of April 29, 1981 in their regular classrooms by their teachers. Each teacher received a bundle of tests with the forms alternated throughout. They were asked to distribute the tests randomly to their classes, read the test administration directions, and, when the hour-long testing period was over, to collect the tests. The investigator then collected all the used and unused tests from the schools. Teachers were also asked not to alter their teaching plans because of the test material. The directions to test administrators are reproduced in Appendix B.

Two weeks later the tests were readministered. In order to ensure that students received the form of the test alternate to the one received on the first occasion, their names were affixed to the proper form before the tests were sent to the participating teachers. Teachers once again read the test directions, collected the tests, and returned all the papers. Two of the nine classes participating postponed the second test administration until the start of the third week in order to accommodate a school play. The classes in the other schools all wrote the tests in the middle of the week. During the week of the second test administration, the principals provided information regarding students' gender and Quantitative IQ score.

Data Analysis

The tests were hand scored by the investigator and checked by a research assistant. The test answer key was constructed by the investigator. On those items in constructed-response format where more than one answer was acceptable, each such response was considered correct. The scores obtained by the investigator and research assistant were in 100% agreement for both forms of the test.

The test data and student data was subsequently entered into a computer file at the U.B.C. Computing Centre. The file was then checked, the errors were corrected, and the file checked once more. There were no errors discovered.

Test Analyses

An item analysis of the test data was performed using the computer program LERTAP (Nelson, 1974). Descriptive statistics were generated by the programs BMDP2D and BMDP2V (Dixon and Brown, 1979).

Preliminary Analyses

The interpretation of an analysis of variance becomes very complicated for large numbers of factors. This is particularly the case if some of the factors are nuisance factors; that is, variables of no particular interest but which must be included because they contribute significantly to the overall variance. The order of test administration, that is, multiple-choice form written first, or constructed-response form written first; and the class in which a student is enrolled are two such variables. Therefore a 2x9x2 (order-by-class-by-item difficulty) fixed effects analysis of variance was performed using the computer program BMDP2V. Order of test administration and class membership were considered as grouping variables and item difficulty was treated as a trial variable with two observations. The difficulty factor was included to add more precision to the analyses.

Six of these analyses were performed, one for each test format in each content domain. A significance level of 0.05 was chosen and 213 cases, that is, all students who completed both forms of the test, were included in the analyses. The summary ANOVA tables for these analyses are contained in Appendix C.

Order of Administration

The results of the analyses showed that order of administration affected the scores of only two of the six subtests. Order of administration did not affect the scores on the multiple-choice subtests in any of the domains but did affect the constructed-response scores in the Application and Algebra domains. In both cases the constructed-response scores for those students who wrote that form of the test second were significantly higher ($p < .05$) than the scores of those students who wrote the constructed-response form first. There were no significant first order interactions involving order of administration and only one significant second order interaction involving order of administration.

Order of test administration had a very limited effect on the overall test scores. In addition, it was not a factor of great interest in the study. Therefore order of administration was eliminated as a factor in further analyses.

Class

The results of the analyses revealed that in three cases of the six there were significant differences ($p < .05$) in subtest scores among classes. Therefore the raw scores were transformed in order to remove the class effects and yet retain all other information. To achieve this, the raw scores were standardized within each class and content domain.

This transformation was performed in the following manner. Within each of the three content domains there are four subtests; two difficulty levels in each of two formats. The mean and standard deviation of the total of the four subtest scores within each class were used to transform the individual raw scores over the four subtests to mean zero and standard deviation one. In order to check the effect of this procedure, the three-way analyses of variance involving class, order, and item difficulty were repeated using the standard scores. As expected, the analyses showed no class effects. All effects involving factors other than class were similar to those computed when raw score data were used. Thus, standard scores were used in all further analyses of variance. A summary of cell means and standard deviations for these standard scores are contained in Appendix D.

Difficulty

As expected, the difficulty of the items caused scores to differ significantly. This factor was retained in subsequent analyses.

Final Analyses

To examine the effects of gender, ability, item format, and item difficulty, a 2x3x2x2 (gender-by-ability-by-format-by-difficulty) analysis of variance procedure was performed using BMDP2V. The data from the 191 students for whom complete information was available were used in this analysis. The students missing ability measures were evenly distributed among classes. Therefore it is reasonable to expect that the class means of zero and standard deviations of one achieved by the transformation of raw scores were not seriously affected by deleting the data from the 22 students for whom ability measures were not available. Gender and three levels of ability were considered as grouping factors. Item format and item difficulty were treated as trial factors each with two levels. A level of significance of 0.05 was chosen. Post hoc comparisons were made using Scheffe's procedure (Kirk, 1968) with a significance level of 0.10. Scheffe (1959, p. 71) suggests this level of significance as appropriate for testing contrasts of this nature, as this test of conservative one. Ferguson (1981, pp. 308-309) also recommends a more liberal level of signficance.

Although the effect of a correction for chance performed on multiple-choice data and the predictive power of multiple-choice scores to constructed-response scores are not central to the research questions, some analyses were performed with regard to these issues. The data from the 191 completed cases were then rescored applying the traditional correction for chance formula. The chance-corrected data were standardized within class using the same procedure as for the uncorrected data and subjected to a 2x3x2x2 (gender-by-ability-by-format-by-difficulty) analysis of variance. The significance level of 0.05 was chosen.. Post-hoc comparisons were made using Scheffe's procedure with a significance level of 0.10.

To estimate the ability of multiple-choice scores to predict constructed-response scores, a simple linear regression analysis was performed. Before the regression analysis was done, however, the raw score data in each content domain were subjected to a 9x2x2x2 (class-by-order-by-difficulty-by-format) fixed effects analysis of variance to check for format-by-class interactions. At the 0.05 level of significance there were no format-by-class interactions using 191 complete cases. The raw scores in each content domain were subsequently employed in a simple linear regression analysis performed using the computer program BMDP1R. Constructed-response score was treated as the dependent variable and multiple-choice score was treated as the independent variable.

Chapter 4

RESULTS

The results of the study are reported in the following order. First, there is a description of the sample, then the results of the descriptive analysis of the test scores and the subtest characteristics are given. Finally, the effects of gender, ability, item format, and item difficulty are presented; and the results of regression analyses performed on the raw scores are discussed.

Description of the Sample

The sample drawn for this study consisted of 237 Grade 7 students enrolled in nine classes in five elementary schools in a suburban-rural school district. Of the total number of students, 24 failed to write one or both forms of the tests due to absence on the testing dates. None of the data from these 24 students were used in any of the analyses. There were also 22 students for whom ability measures were unavailable. The data from these students were included only in the descriptive analyses of the test items and the inferential analysis of the effect of class membership and order of test administration. There were, then 191 students, 88 boys and 103 girls, from whom complete sets of data were obtained.

The ability measures for the 191 complete cases were IQ scores from the Canadian Cognitive Abilities Quantitative Battery (Thorndike, Hagen, and Wright, 1974). The mean IQ score was 102.6 with a standard deviation of 13.9. The sample was partitioned into low, average, and high ability groups of roughly equal size. Students with Quantitative IQ scores of 96 or less were considered to be low ability students; students with scores greater than 96 but less than or equal to 107 were considered to be of average ability; and students with scores higher than 107 were considered to be high ability students. Table 4.1 contains the distribution of students by ability and gender.

Table 4.1

Distribution of Subjects by Gender and Ability

Ability	Male	Female	Total
High	33	32	65
Average	24	40	64
Low	31	31	62
Total	88	103	191

Descriptive Analysis of the Test Scores

There are four sets of information regarding item difficulty and subtest reliability. Difficulty indices for the multiple-choice items are available from the pilot testing in Grade 8 for the 1981 Assessment (Robitaille, 1981), from the pilot testing performed in Grades 7 and 8 as part of the present study, and from the main study itself. There is test reliability information from three sources, the Grades 7 and 8 pilots and from the main study. The results of the Grade 8 pilot indicated that for Grade 8 subjects, the items were too easy. Therefore subjects in Grade 7 were selected for the study.

Table 4.2 contains the average *p*-values for the items on each subtest from the Assessment pilot, the Grade 7 pilot and the main study. The items were less difficult than one would expect on the basis of the Assessment pilot data. The high difficulty multiple-choice items were chosen so as to have an average *p*-value of approximately 0.4. In the main study the *p*-values for the Computation and Application domains were 0.529 and 0.566 respectively. The high difficulty Algebra multiple-choice items, however had an average *p*-value of 0.388. Similarly, the *p*-values of the low difficulty multiple-choice items were expected to be

Table 4.2

Mean Item Difficulty

Subtest	Assessment	Grade 7 Pilot		Study	
	Pilot				
	M-C	M-C	C-R	M-C	C-R
	n=240	n=28	n=29	n=213	n=213
Computation					
High Difficulty	.404	.591	.483	.529	.456
	(.058)	(.148)	(.199)	(.121)	(.170)
Low Difficulty	.688	.842	.827	.808	.716
	(.036)	(.074)	(.113)	(.076)	(.100)
Application					
High Difficulty	.429	.638	.468	.566	.413
	(.044)	(.136)	(.124)	(.079)	(.104)
Low Difficulty	.710	.893	.799	.832	.749
	(.073)	(.077)	(.098)	(.087)	(.156)
Algebra					
High Difficulty	.402	.571	.320	.388	.232
	(.053)	(.196)	(.235)	(.104)	(.172)
Low Difficulty	.702	.842	.630	.717	.521
	(.033)	(.068)	(.235)	(.099)	(.258)

Note. Each subtest contained 7 items.
The numbers in parentheses are the standard deviations of the *p*-values.

approximately 0.7 but they had computed values of 0.808, 0.832, and 0.717 for the three content domains Computation, Application, and Algebra. Nonetheless, although they were easier than anticipated the items were partitioned into two distinct difficulty levels in each content domain.

Test reliabilities are influenced by a number of factors including item difficulties and length of test. Each of the subtests in this study contained only a small number of items (seven), and a restricted range of difficulty. As a result it is not surprising that the subtest reliabilities computed using Hoyt's ANOVA procedure are not high. They range from a low of 0.47 to a high of 0.76. These reliabilities are found in Table 4.3.

Inferential Analyses

Two sets of inferential analyses were performed on the data. The preliminary analyses were done to determine whether or not variables which were considered to be nuisance variables could be safely eliminated from the final analyses. These preliminary analyses were discussed in Chapter 3. The final analyses were performed using the variables of greatest interest in the study. The final analyses are discussed below.

Table 4.3

Subtest Reliabilities, Means, Standard Deviations, and Standard Errors of Measurement

Subtest	Mean Score	Hoyt Reliability	Standard Error
Computation high difficulty			
Multiple-choice	3.71 (1.77)	0.55	1.10
Constructed-response	3.19 (1.70)	0.53	1.08
Computation low difficulty			
Multiple-choice	5.66 (1.38)	0.52	0.88
Constructed-response	5.01 (1.52)	0.47	1.02
Application high difficulty			
Multiple-choice	3.96 (1.75)	0.53	1.12
Constructed-response	2.90 (1.80)	0.58	1.09
Application low difficulty			
Multiple-choice	5.83 (1.42)	0.62	0.81
Constructed-response	5.24 (1.59)	0.62	0.90
Algebra high difficulty			
Multiple-choice	2.72 (1.96)	0.68	1.03
Constructed-response	1.62 (1.60)	0.68	0.84
Algebra low difficulty			
Multiple-choice	5.02 (1.75)	0.65	0.96
Constructed-response	3.65 (1.98)	0.76	0.89

Note. Each subtest contained 7 items.
Two hundred thirteen subjects wrote each subtest.
The numbers in parentheses are standard deviations.

Gender, Ability, Item Format, and Item Difficulty

The effects due to gender, ability, item format, and item difficulty were examined in each content domain using a 2x3x2x2 (gender-by-ability-by-format-by-difficulty) analysis of variance with repeated measures. Gender and ability were considered grouping factors and item format and item difficulty were trial factors. Tables 4.10 to 4.12 show the summary ANOVA tables for the analyses in each of the content domains. The effects due to each variable are presented separately below and then the significant interactions are presented. A significance level of 0.05 was chosen for each of the omnibus *F*'s.

Gender

An examination of the summary ANOVA tables found in tables 4.4 to 4.6 indicates that there was a significant effect due to gender in only one of the domains, Application. In this case, the mean score obtained by males was significantly higher than that obtained by females.

Ability

In each of the three content domains, the analyses of the data revealed that there were significant effects due to

Table 4.4

Summary Analysis of Variance

Gender, Ability, Item Format, and Item Difficulty

Computation Domain

Source of variance	Degrees of Freedom	Mean Square	F
Mean	1	1.078	0.49
Gender (G)	1	3.178	1.43
Ability (A)	2	121.791	54.84*
G x A	2	3.384	1.52
Error	185	2.221	
Format (F)	1	49.568	70.61*
F x G	1	0.206	0.29
F x A	2	0.044	0.06
F x G x A	2	0.890	1.27
Error	185	0.702	
Difficulty (D)	1	502.638	380.96*
D x G	1	0.198	0.15
D x A	2	5.253	3.98*
D x G x A	2	3.325	2.52
Error	185	1.319	
F x D	1	0.034	0.06
F x D x G	1	1.628	2.76
F x D x A	2	0.632	1.07
F x D x G x A	2	0.753	1.27
Error	185	0.591	

*p < .05

Table 4.5
Summary Analysis of Variance
Gender, Ability, Item Format, and Item Difficulty
Application Domain

Source of variance	Degrees of Freedom	Mean Square	F
Mean	1	0.753	0.33
Gender (G)	1	11.440	4.94*
Ability (A)	2	138.500	59.86*
G x A	2	4.023	1.74
Error	185	2.314	
Format (F)	1	71.234	95.72*
F x G	1	2.891	3.88
F x A	2	2.133	2.87
F x G x A	2	1.252	1.68
Error	185	0.744	
Difficulty (D)	1	492.124	484.47*
D x G	1	4.154	4.09*
D x A	2	3.262	3.21*
D x G x A	2	1.316	1.30
Error	185	1.016	
F x D	1	7.321	13.90*
F x D x G	1	0.113	0.21
F x D x A	2	0.270	0.51
F x D x G x A	2	0.079	0.15
Error	185	0.527	

*p < .05

Table 4.6

Summary Analysis of Variance

Gender, Ability, Item Format, and Item Difficulty

Algebra Domain

Source of variance	Degrees of Freedom	Mean Square	F
Mean	1	2.416	1.03
Gender (G)	1	0.455	0.19
Ability (A)	2	133.618	56.70*
G x A	2	0.217	0.09
Error	185	2.357	
Format (F)	1	122.306	225.05*
F x G	1	0.262	0.48
F x A	2	0.157	0.29
F x G x A	2	0.745	1.37
Error	185	0.543	
Difficulty (D)	1	412.578	491.77*
D x G	1	0.061	0.07
D x A	2	2.696	3.21*
D x G x A	2	0.047	0.06
Error	185	0.839	
F x D	1	1.073	2.91
F x D x G	1	0.041	0.11
F x D x A	2	4.555	12.37*
F x D x G x A	2	0.184	0.50
Error	185	0.368	

*p < .05

ability. The mean scores of the students in each ability level were compared using Scheffe's procedure with a significance level of 0.10. It was found that the means of the total scores in each content domain were ordered strictly according to ability level. Students of high ability scored significantly higher than students of average ability, who, in turn, scored significantly higher than students of low ability.

Item Format

In each domain, the format had a significant effect on scores. In each case multiple-choice scores were higher than constructed-response scores.

Item Difficulty

The analyses showed that there was an effect due to item difficulty. In each domain, the mean scores on items of low difficulty were significantly greater than the mean scores on items of high difficulty.

First Order Interactions

The summary ANOVA tables show five significant first-order interactions: item difficulty by gender in the Application domain, item difficulty by ability in all three content domains, and item format by item difficulty in the Application domain. Many contrasts can be formed to test interactions. For the purposes of this study, tetrad differences were considered to be the only contrasts of interest. The tetrad differences for each of the significant interactions were analysed using Scheffe's procedure with an alpha level of 0.10. The results of the analyses of each of the first order interactions is discussed below.

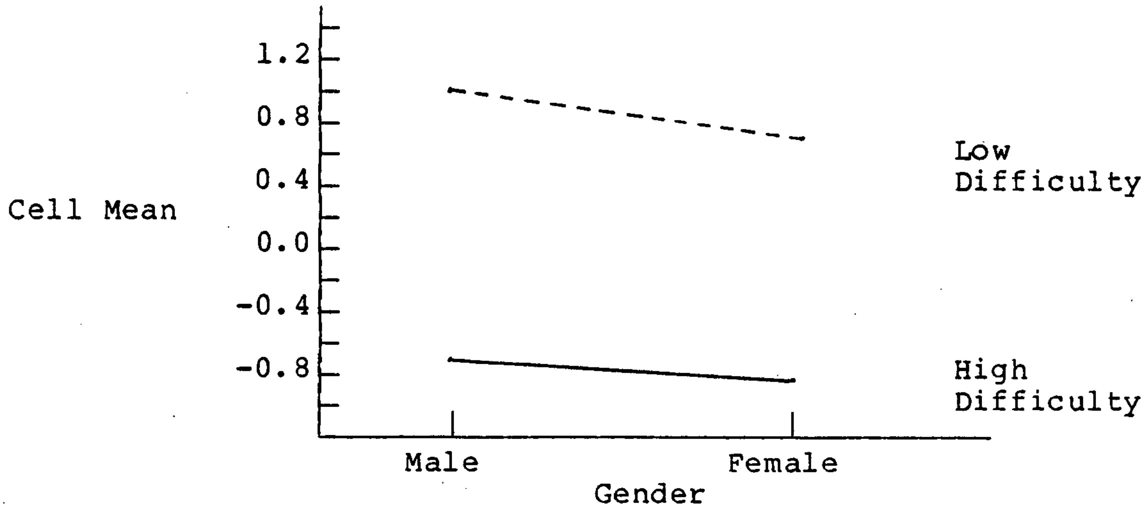
Difficulty by Gender:- The difficulty by gender interaction resulted in a significant omnibus F-ratio only in the Application domain. An analysis of the tetrad differences, however, failed to show significant differences. Figure 4.1 contains a graph of cell means versus gender for each of the item difficulty levels. The failure of the test of significance on the tetrad differences indicates that the change in performance of males on high difficulty items when compared to their performance on low difficulty items is not different than the corresponding change in performance for females. The significant omnibus F-ratio implies that one could construct a complex comparison of the difficulty-by-gender cell means which would be statistically significant, however such a comparison would not be helpful in answering the experimental questions.

Figure 4.1

Plot of Cell Mean versus Gender

for Items of High and Low Difficulty

Application Domain



Difficulty by Ability:- The difficulty by ability interaction resulted in a significant omnibus F in all three domains. Figure 4.2 contains graphs of cell means versus item difficulty for the Computation domain. Similar graphs are found in Figures 4.3 and 4.4 for the Application and Algebra domains respectively.

The tetrad differences for each of these interactions were analysed and the null hypothesis was not rejected in any case. It appears that the differences between performance on high difficulty and low difficulty items do not vary signficantly at the 0.10 level among ability levels for any of the content domains.

Figure 4.2

Plot of Cell Means versus Item Difficulty by Ability Level

Computation Domain

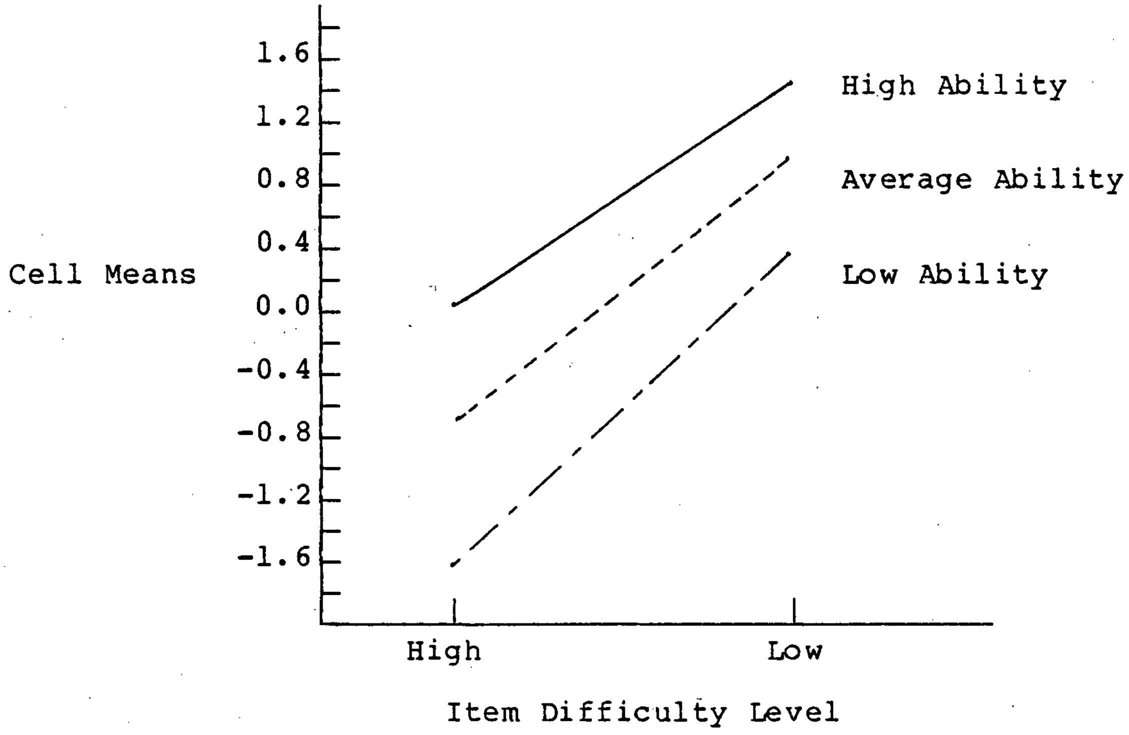


Figure 4.3

Plot of Cell Means versus Item Difficulty by Ability Level

Application Domain

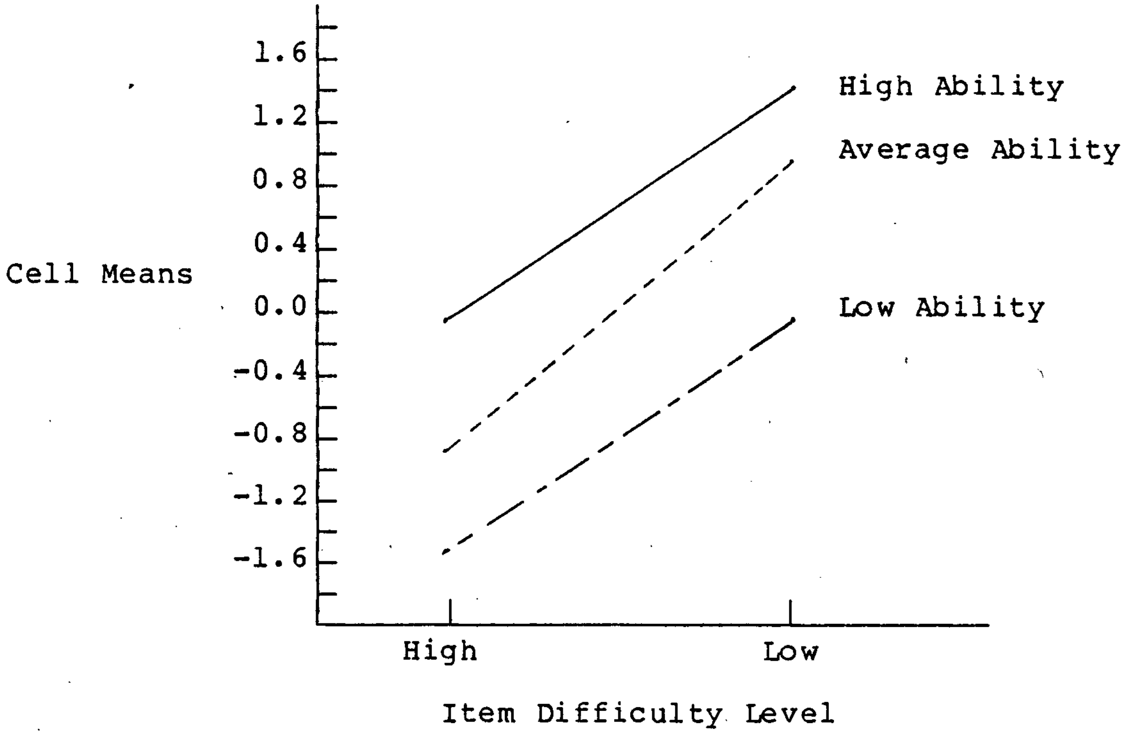
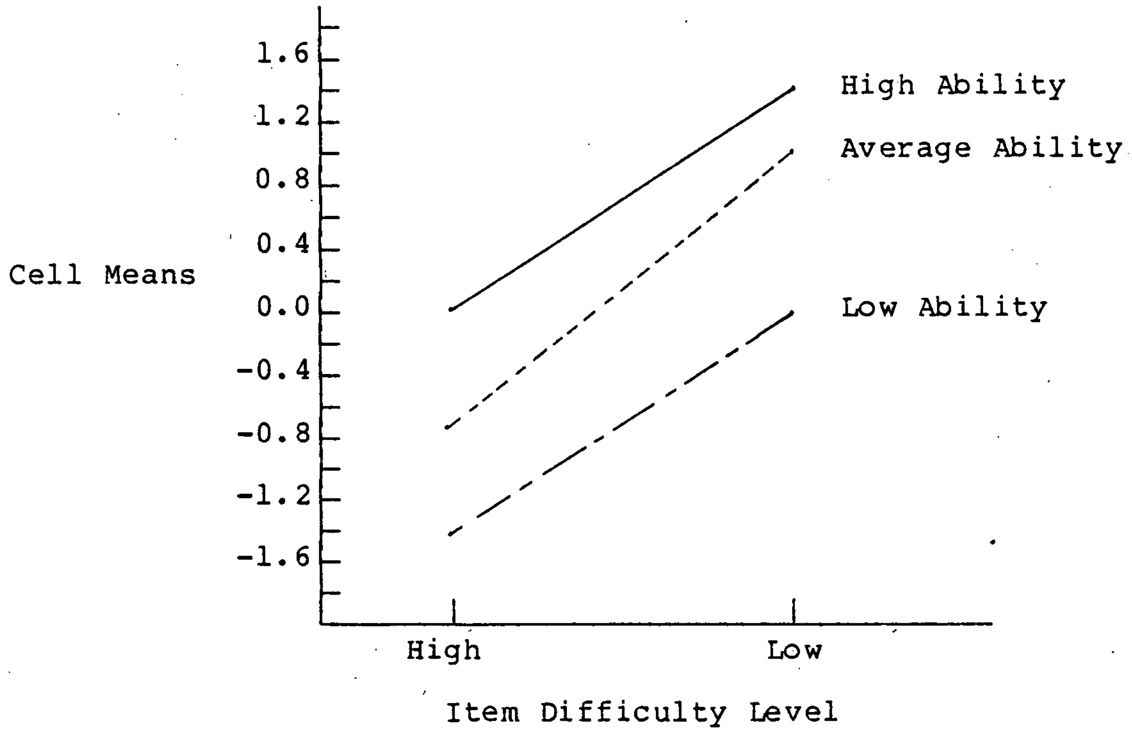


Figure 4.4

Plot of Cell Means versus Item Difficulty by Ability Level

Algebra Domain

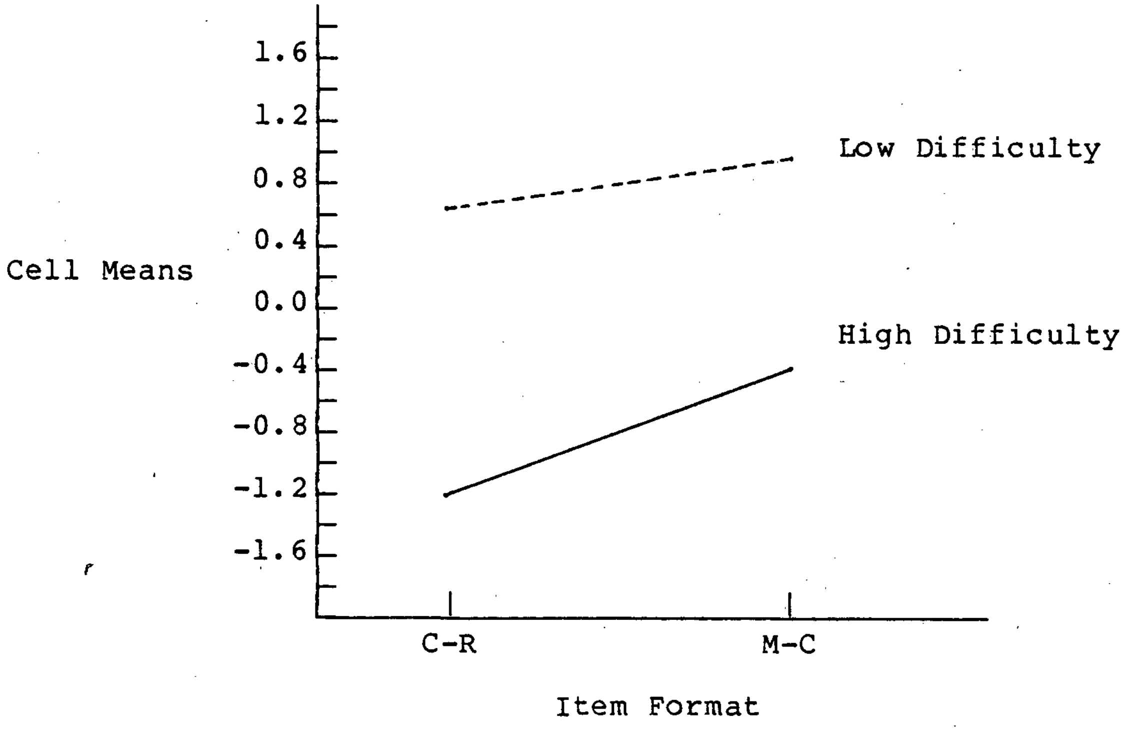


Format by Difficulty:- The format by difficulty interaction resulted in a significant omnibus F statistic in the Application domain. A graph of cell means versus item format for two levels of difficulty is found in Figure 4.5.

Figure 4.5

Plot of Cell Means versus Item Format by Item Difficulty

Application Domain



Scheffe's test showed the tetrad difference to be significant (p < 0.10). That is, the difference in scores on items of high difficulty in multiple-choice and constructed-response formats was significantly different than the difference in scores on items of low difficulty in the two formats. There is a greater difference in achievement between formats for difficult items than for easy items.

Second Order Interactions

The summary ANOVA tables show one significant omnibus _F_ for a second order interaction. That interaction, item format by item difficulty by ability, was found in the Algebra domain. Figure 4.6 shows three graphs. These graphs plot cell mean versus item format at each of two levels of difficulty for the three ability levels. An analysis of the tetrad differences shows that while the interactions of item format by item difficulty for average and high ability students are not significantly different from one another, both of these are significantly different from the interaction of item format by item difficulty for low ability students.

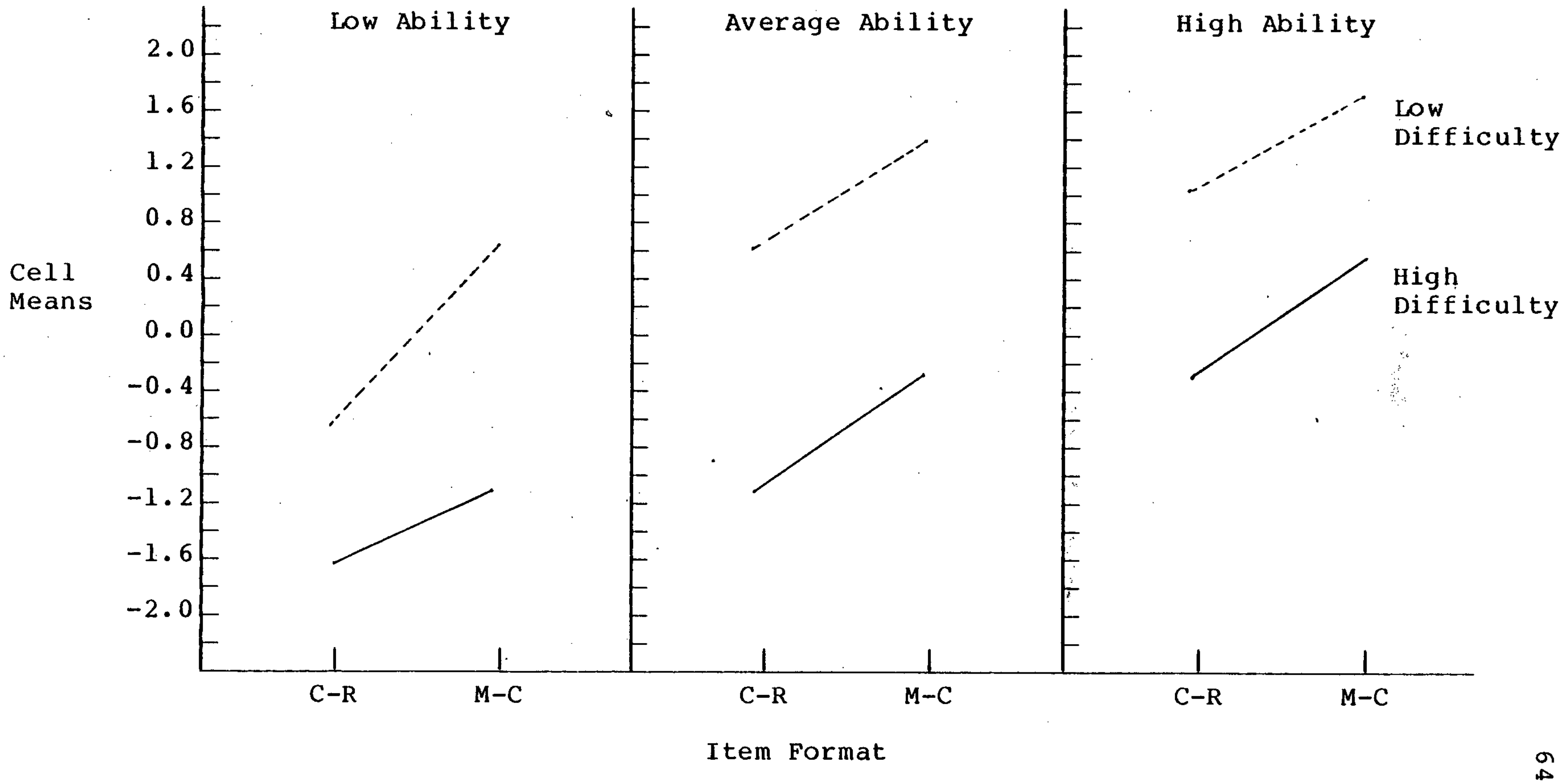
Correction for Guessing

The multiple-choice scores were transformed using the traditional correction for guessing and then standardized within the class and subjected to the same set of 2x3x2x2 (gender-by-ability-by-format-by-difficulty) analyses of variance as the uncorrected scores (Appendix D contains a

Figure 4.6

Plot of Cell Mean versus Item Format by Item Difficulty for Three Ability Levels

Algebra Domain



summary of the cell means and standard deviations for both corrected and uncorrected scores). The results of these analyses are summarized in Tables 4.7 to 4.9. These results followed a pattern very similar to those findings of the analyses performed on the scores not corrected for chance. The findings are summarized in Figure 4.9. In particular the main effects were identical for both sets of data except for the effect of format. The format effect (multiple-choice score greater than constructed-response score) was significant ($p < .05$) in the Application and Algebra domains but not significant in the Computation domain.

The patterns of significant interactive effects were also very similar. For the corrected data there were significant interactions for difficulty-by-gender in the Application domain but no significant tetrad differences ($p < .10$) were found. Similarly for the significant difficulty-by-ability interactions in the Computation and Algebra domains no significant tetrad differences were found.

There were significant format-by-difficulty interactions in the Computation and Algebra domains. Significant tetrad differences were found in these interactions. These tetrad differences indicate that the difference between scores on the low difficulty and high difficulty subtests in multiple-choice format is larger than the difference between

Table 4.7
Summary Analysis of Variance
Gender, Ability, Item Format, and Item Difficulty
Computation Domain
Multiple-Choice Scores Corrected for Guessing

Source of variance	Degrees of Freedom	Mean Square	F
Mean	1	1.048	0.47
Gender (G)	1	2.979	1.33
Ability (A)	2	123.043	55.07*
G x A	2	3.647	1.63
Error	185	2.234	
Format (F)	1	1.292	1.77
F x G	1	0.045	0.06
F x A	2	1.438	1.97
F x G x A	2	1.281	1.75
Error	185	0.731	
Difficulty (D)	1	519.856	378.84*
D x G	1	0.371	0.27
D x A	2	5.681	4.14*
D x G x A	2	3.550	2.59
Error	185	1.372	
F x D	1	5.949	9.75*
F x D x G	1	2.007	3.29
F x D x A	2	1.123	1.84
F x D x G x A	2	1.059	1.74
Error	185	0.610	

*p < .05

Table 4.8

Summary Analysis of Variance

Gender, Ability, Item Format, and Item Difficulty

Application Domain

Multiple-Choice Scores Corrected for Guessing

Source of variance	Degrees of Freedom	Mean Square	F
Mean	1	0.570	0.25
Gender (G)	1	10.330	4.44*
Ability (A)	2	140.988	60.62*
G x A	2	4.262	1.83
Error	185	2.326	
Format (F)	1	13.831	18.38*
F x G	1	2.286	3.04
F x A	2	1.929	2.56
F x G x A	2	0.824	1.09
Error	185	0.752	
Difficulty (D)	1	503.352	475.12*
D x G	1	4.594	4.34*
D x A	2	3.148	2.97
D x G x A	2	0.953	0.90
Error	185	1.059	
F x D	1	0.300	0.52
F x D x G	1	0.317	0.55
F x D x A	2	0.162	0.28
F x D x G x A	2	0.091	0.16
Error	185	0.575	

*p < .05

Table 4.9
Summary Analysis of Variance
Gender, Ability, Item Format, and Item Difficulty
Algebra Domain
Multiple-Choice Scores Corrected for Guessing

Source of variance	Degrees of Freedom	Mean Square	F
Mean	1	2.257	0.95
Gender (G)	1	0.324	0.14
Ability (A)	2	135.774	57.23*
G x A	2	0.327	0.14
Error	185	2.372	
Format (F)	1	24.969	42.87*
F x G	1	0.184	0.32
F x A	2	0.893	1.53
F x G x A	2	0.751	1.29
Error	185	0.582	
Difficulty (D)	1	426.849	481.29*
D x G	1	0.009	0.01
D x A	2	2.766	3.12*
D x G x A	2	0.082	0.09
Error	185	0.887	
F x D	1	8.433	20.90*
F x D x G	1	0.006	0.02
F x D x A	2	4.952	12.27*
F x D x G x A	2	0.201	0.50
Error	185	0.404	

*p < .05

scores on the low difficulty and high difficulty subtests in constructed-response formats. These interactions are plotted in Figures 4.7 and 4.8.

As in the case of the uncorrected data, there was one significant format-by-difficulty-by-ability interaction in the Algebra domain. The character of this interaction was identical for both corrected and uncorrected data; the same set of tetrad differences were significant for both.

Regression Analyses

A set of simple linear regression analyses were performed on the raw score data. In these analyses, the constructed-response scores were treated as the dependent variables and the multiple-choice scores were considered to be the independent variables.

A series of analyses of variance, one analysis for each content domain preceded the regression analyses. These were done to ensure that there were no significant class by format interactions in the raw score data which would confound the regressions. No such interactions were found.

The regression weights, correlations between the scores in each format, and the standard errors are found in Table 4.10. In each of the content domains, the correlations were

Figure 4.7

Plot of Cell Means versus Item Format by Item Difficulty

Computation Domain

Chance-Corrected Data

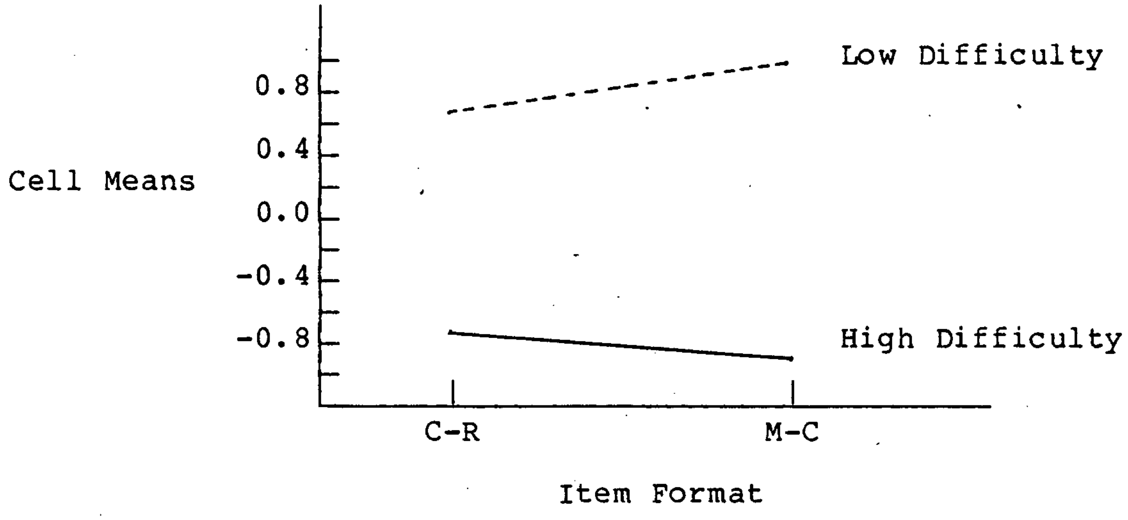


Figure 4.8

Plot of Cell Means versus Item Format by Item Difficulty

Algebra Domain

Chance-Corrected Data

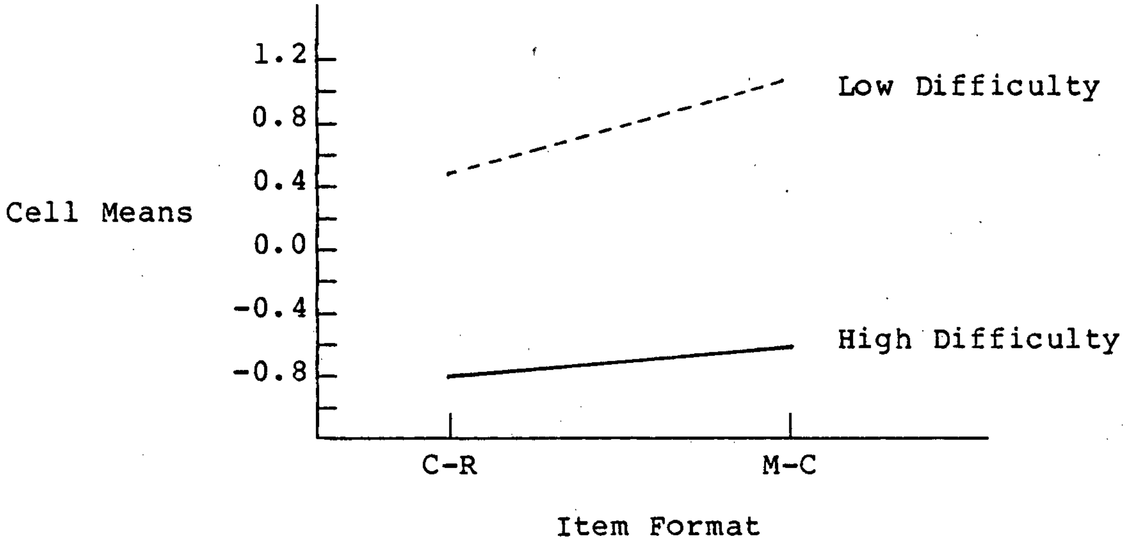


Figure 4.9

Summary of Main Effects and Interactions

Main Effects

	Computation	Application	Algebra
Gender		✓	
Ability	✓	✓	✓
Format	✓	✓	✓
Difficulty	✓	✓	✓

Scores Corrected for Guessing

	Computation	Application	Algebra
Gender		✓	
Ability	✓	✓	✓
Format		✓	✓
Difficulty	✓	✓	✓

Interactions

	Computation	Application	Algebra
Difficulty X Gender		✓	
Difficulty X Ability	✓	✓	✓
Format X Difficulty		*	
Format X Diff X Ability			*

Scores Corrected for Guessing

	Computation	Application	Algebra
Difficulty X Gender		✓	
Difficulty X Ability	✓		✓
Format X Difficulty	*		*
Format X Diff X Ability			*

✓: Significant Omnibus F
*: Significant Tetrad Differences

greater than 0.7 which indicate that there is a strong relationship between the scores in each format. The multiple R-squared statistics show that in each domain, approximately half of the variance in the open-ended scores can be predicted by the multiple-choice scores.

Table 4.10

Regression Weights and Intercepts, and Correlation Coefficients for C-R Scores Regressed on M-C Scores

Test	Weight	Intercept	Standard Error	Correlation
Computation	0.717	1.497	0.050	0.721
Application	0.807	0.274	0.056	0.726
Algebra	0.768	-0.604	0.046	0.772

Chapter 5

SUMMARY, CONCLUSIONS, AND IMPLICATIONS

This study was undertaken to determine the effect of item format on students' scores on a mathematics achievement test. A total of 191 Grade 7 students who wrote the achievement test in two formats, multiple-choice and constructed-response, one on each of two occasions, provided the data for the study. These data were analysed using an analyses of variance procedure with repeated measures. Other factors besides item format were considered. These factors included gender, ability, and item difficulty. The results of the analyses shed some light on the relationship between the scores obtained on multiple-choice and constructed-response mathematics achievement tests.

In the paragraphs which follow, a summary of the findings and conclusions based on those findings are presented. Implications of the results are then discussed and then suggestions for future research are presented.

Summary

Item Format

Test results showed that item format had a significant effect on the scores obtained by students on a mathematics achievement test. In each of the three content domains; Computation, Application, and Algebra; scores on the multiple-choice form of the test were higher than on the constructed-response form. It would therefore be unwise to consider the multiple-choice and constructed-response scores as interchangeable measures. However, although the scores are of different magnitude, it may be that with a suitable change of scale, one score can be transformed into the other.

In order to feel comfortable about such a transformation, however, it would be necessary to consider any possible confounding effects due to other variables. Previous research has shown that multiple-choice test scores may be affected by response sets, particularly guessing. It has also been shown that response sets are related to personality characteristics. The equivalence of scores depends not only on the transformation of the scores, but also upon the assumption that other factors do not interact with item format to produce unique effects.

The findings regarding other variables considered in the study are discussed below. Particular attention is given to describing the nature of any interactions observed.

Class

It was found that there were differences in achievement among the nine classes sampled. Although one might have hoped for a consistent level of performance among classes at the same grade level, it is not surprising that such differences exist. More important to the questions of the study, however, is the absence of significant format by class interactions. This suggests that although the levels of performance among classrooms are significantly different, none of those differences can be attributed to the effect of item format.

Gender

In the Application domain males' scores were significantly higher than females'. Sex-related differences in the other two domains were not significant. These findings are in agreement with recent studies (Robitaille and Sherrill, 1977; Robitaille, 1981; Swafford, 1980) which show that if differences in achievement exist between males and females, males tend to outperform females only on those items which require the application of higher cognitive skills.

There were no significant interactions between format and gender. This indicates that both males and females tend to respond to multiple-choice tests in the same way and respond to constructed-response items in the same way. Therefore, if one wishes to use multiple-choice scores as an indication of probable score on the same test in constructed-response format, there is no need to make an adjustment on the basis of a subject's gender.

There was a significant omnibus F obtained for the item difficulty by gender interaction in the Application domain. The interaction, however, was such that there were no significant contrasts among the means which are relevant to the questions of the study.

Ability

As expected, in all content domains high ability students scored significantly higher than average ability students who, in turn, scored significantly higher than low ability students. There were also significant item difficulty by ability interactions in all three domains. However, there were no contrasts among the means which produced significant results and which were meaningful in terms of the experimental questions. Therefore, differences in scores between high and low difficulty items did not vary according to ability level.

The ability variable did not interact with item format to produce unique effects in any domain. It is reasonable to suspect that low ability students might achieve undeserved higher scores due to guessing. Because there are more items for which low ability students do not know the answer these students have more of an opportunity to guess. Answers guessed correct will then inflate the scores to a higher degree than for students who did not guess. This suspicion was not confirmed by the data: the ability by format interactions were not significant. Although there are score differences related to ability, the scores are not affected by a unique combination of ability and item format. Students of differing abilities are not differentially affected by item format. This finding also supports earlier claims that response sets are relatively independent of ability (Hopkins, 1964).

Item Difficulty

Results show that students' scores were lower on difficult items than on easy items in all content domains. This finding is not of great interest. Of greater interest is the interaction between item format and item difficulty. There was a significant interaction between format and difficulty in the Application domain only.

An analysis of the interaction showed that the difference between multiple-choice and constructed-response scores on high difficulty items was greater than the corresponding difference for low difficulty items.

It is not obvious why this interaction should exist only in the Application domain. It may be that content in the Application domain is familiar enough so that clues provided in the multiple-choice alternatives were enough to elicit a correct response. In contrast, the Computation domain may have been so familiar that the students' responses were not affected by clues. The students knew whether or not they could do the exercise and therefore did not search for clues. In the Algebra domain, the content may have been so unfamiliar that clues were of no help. It may also be that because the Application domain contained items that were applications of mathematics to real situations, students may have been more willing or able to seek reasonable responses from those provided in the list of alternatives.

There was a significant second-order interaction involving item format, item difficulty, and ability in the

Algebra domain. The analysis of this interaction indicated that the differences between high and low difficulty item scores did not vary with format for high and average ability students. However, for students of low ability, the difference between scores on high and low difficulty items in constructed-response format was less than the corresponding differences for students of high and average ability. Similarly, the difference between scores on high and low difficulty items in multiple-choice format was greater than the corresponding differences for high and average ability students. The reader is referred to Figure 4.6.

This pattern of achievement may also be due to the value of clues supplied to the students by the multiple-choice alternatives. For difficult items, the difference between scores over formats for low ability students is less than that of more able students indicating that perhaps the clues did not help students choose a response for an item about which they were unfamiliar. On the other hand, for easy items, the clues may have provided low ability students with enough information to make a reasonable guess.

The reason that format interacts with item difficulty is not clear. However, significant effects due to this interaction are not widespread. Therefore it would be ill-advised to claim that this effect is of major importance.

Conclusions

The results of this study appear to answer the experimental questions. First, students do score higher on a multiple-choice form of a mathematics achievement test than on the same test in constructed-response format. With regard to gender, males significantly outperformed females in only one domain, Application. There are no differential effects due to format related to gender. Similarly, there are no differential effects due to format related to ability or related to the class in which a student is enrolled.

The effects of item difficulty combined with format are not clear. Format and difficulty showed a significant first-order interaction in only one content domain and these two factors were also involved in a second order interaction in another domain. It appears that format and difficulty have a unique effect only when the student has partial knowledge and is able to make use of the clues provided in the alternatives of the multiple-choice questions. This speculation, however, is only weakly supported by the results.

The absence of format interactions is encouraging. This indicates that one may be able to develop a procedure which, when applied to multiple-choice scores, will transform them into the scores which would have been obtained if the subjects had written the test in constructed-response format. This transformation would be such that it would not penalize, or give advantage to, any particular group of students from a population partitioned by class, gender, or ability.

The multiple-choice scores were corrected for chance and then analysed in the same manner as the uncorrected scores. The differences due to test format were eliminated in only the Computation domain. In addition there were format-by-difficulty interactions in the Computation and Algebra domains. It appears that the traditional correction for guessing is not sufficient to make multiple-choice scores equivalent to constructed-response scores.

The constructed-response scores obtained in the study were regressed on the multiple-choice scores. The results of the simple linear regression showed that the constructed-response scores were moderately to strongly correlated with multiple-choice scores. The correlations ranged between 0.72 and 0.77 for the three content domains. This implies that the multiple-choice scores account for between 52% and 60% of the variance in the constructed-response scores. Given that the test reliabilities were somewhat low, there were a small number of items, and there was a restricted range of item difficulties, it may be that on longer, more reliable tests these correlations would be significantly higher.

The regression weights were computed and found to range between 0.72 and 0.81. the intercept values, however ranged between approximately 1.5 and -0.60. Although the slopes of the regression lines are fairly constant, the intercepts are not. This indicates that it is unlikely that there is a global scoring formula which can be applied to all multiple-choice tests to obtain an estimate of the constructed-response score. Rather, the scoring formula for each test may have to be determined empirically.

Implications

The results of the study make it clear that the scores of multiple-choice and constructed-response tests are not interchangeable when making criterion-referenced judgments. However, the results indicate that the two measures are equivalent except for a change of scale. In fact, the change of scale may be a simple linear transformation. This finding is in agreement with that of Rowley and Traub (1977).

Mason's (1979) objection to basing criterion-referenced judgments on multiple-choice data is largely unfounded.

Although the absolute scores are not equivalent, the items in both formats appear to measure the same attributes in the same way. The format apparently does not have a differential effect according to content variables or subject variables. Therefore, after allowing a fixed amount for format differences, the two types of scores are interpretable in the same manner.

These findings make the task of interpreting the results of multiple-choice assessment tests less ambiguous. The claim that multiple-choice and constructed-response tests rank students in essentially the same order (Cronbach, 1970) is confirmed. It also appears that multiple-choice scores are obtained in such a way that a comparison of intervals within those scores may be interpreted in the same way as intervals within scores as though they had been obtained by constructed-response items. Therefore, the results of multiple-choice tests do not need to be considered as simply ordinal information. The scores can be used to compare groups in other than a norm-referenced fashion. The skepticism of Interpretation Panels toward making criterion-referenced judgments, as reported by Mussio and Greer (1980) may then be alleviated.

Limitations of the Study

Although the present study investigated the effect of item format on achievement test scores, the following conditions are limiting factors. Only students enrolled in grade 7 were sampled, and all of those students attended schools in the same suburban/rural school district. The content sampled by the test items was mathematics and did not represent the total Grade 7 mathematics curriculum. The items tested the areas of Computation, Application, and Algebra.

Suggestions for Future Research

Results of this study have cast some light on the relationship between multiple-choice and constructed-response test scores. A number of questions remain unanswered, however. These may be pursued by further research.

In this study constructed-response scores were regressed on multiple-choice scores to obtain estimates of the correlation between the two types of scores. At first sight, the consistency of the regression weights suggests that there may be a simple linear transformation which would change one score into the other. It may be possible to determine this transformation empirically. Studies directed toward determining an empirical scoring formula should be pursued.

In the study there was no attempt made to analyse the discrepancies between the two types of scores. However, the multiple-choice scores were subjected to the traditional correction for guessing procedure and reanalysed. The correction for guessing did not remove the format effect. The usefulness of this procedure should be investigated empirically, perhaps to examine the predictive ability of corrected scores. Such an examination might also provide information about the effect of omitted items on constructed-response score prediction.

In the discussion of the results there was mention made of the effect of partial knowledge. Partial knowledge is the body of facts and understandings which, although it does not allow the respondent to construct or choose the correct answer with certainty, enables him or her to eliminate unlikely responses or aids in constructing an acceptable response. There are other multiple-choice formats which attempt to measure the extent of the subject's partial knowledge. An examination of the effect of partial knowledge on the predictive ability of multiple-choice test scores may shed some light on the item format by item difficulty interactions found in this study.

REFERENCES

Bracht, G. H. and Hopkins, K. D. Comparative validities of essay objective tests. Research paper No. 20. Boulder: University of Colorado, Laboratory of Educational Research, 1968.

Bracht, G. H. and Hopkins, K. D. The communality of essay and objective tests of academic achievement. Educational and Psychological Measurement, 1970, 30, 359-364.

British Columbia Ministry of Education. Mathematics curriculum guide years one to twelve. Victoria, B.C.: 1978.

Burton, N. W. An investigation of item-scoring formulas which take into account random guessing, partial information, and misinformation. Unpublished doctoral dissertation, University of Colorado, 1972.

Cronbach, L. J. Response sets and test validity. Educational and Psychological Measurement, 1946, 6, 475-494.

Cronbach, L. J. Essentials of psychological testing (3rd ed.). New York: Harper and Row, 1970.

Dixon, W.J. and Brown, M.B. Biomedical Computer Programs, P-Series. Berkeley: University of California Press, 1979.

Ebel, R. L. Blind guessing on objective achievement tests. Journal of Educational Measurement, 1968, 5, 321-325.

Ebel, R. L. *Essentials of educational measurement* (3rd ed.). Englewood Cliffs, N.J.: Prentice-Hall, 1979.

Erickson, G., Erickson, L., and Haggerty, S. *Gender and mathematics/science education in elementary and secondary schools*. Discussion paper 08/80. Victoria, B.C.: Ministry of Education, 1980.

Fennema, E., Wolleat, P. L., Pedro, J. D., and Becker, A. D. Increasing women's participation in mathematics: An intervention study. *Journal for Research in Mathematics Education*, 1981, *12*, 3-14.

Ferguson, G.H. *Statistical analysis in psychology and education* (5th ed.). New York: McGraw-Hill, 1981.

Gronlund, N. E. *Constructing achievement tests*. Englewood Cliffs, New Jersey: Prentice-Hall, 1968.

Hopkins, K. D. Extrinsic reliability: Estimating and attenuating variance from response sets, chance, and other irrelevant sources. *Educational and Psychological Measurement*, 1964, *24*, 271-281.

Kinney, L. B. and Eurich, A. C. A summary of investigations comparing different types of tests. *School and Society*, 1932, *36*, 540-544.

Kirk, R. E. *Experimental design: Procedures for the behavioral sciences*. Belmont, California: Wadsworth, 1968.

Klassen, W., Dukowski, L., and deGroot, I. Item preparation for the 1981 B.C. Mathematics Assessment. Vector, Winter, 1981, 22 (2), 22-25.

Lord, F. M. Formula scoring and validity. Educational and Psychological Measurement, 1963, 23 (4), 663-672.

Lord, F. M. Formula scoring and number right scoring. Journal of Educational Measurement, 1975, 12, 7-11.

Lord, F. M. and Novick, M. R. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.

Mason, G. P. Test purpose and item type. Canadian Journal of Education, 1979, 4 (4), 8-13.

Mussio, J. J. and Greer, R. N. The British Columbia Assessment Program: An overview. Canadian Journal of Education, 5 (4), 1980, 22-40.

Nelson, L. R. Guide to LERTAP Use and Interpretation. Dunedin, New Zealand: University of Otago, 1974.

Rabinowitz, J. C., Mandler, G., and Patterson, K. Determinants of recognition and recall: Accessibility and generation. Experimental Psychology: General, 1977, 106, 302-329.

Robitaille, D. (Ed.). The 1981 B.C. mathematics assessment: General report. Victoria, B.C.: Ministry of Education, 1981.

Robitaille, D. and Sherrill, J. British Columbia Mathematics Assessment: Summary Report. Victoria, B.C.: Ministry of Education, 1977.

Rowley, G. L. and Traub, R. E. Formula scoring, number-right scoring, and test taking strategy. Journal of Educational Measurement, 1977, 14, 15-22.

Scheffe, H. A. The analysis of variance. New York: John Wiley and Sons, 1959.

Sherriffs, A. C. and Boomer, D. S. Who is penalized by the penalty for guessing? Journal of Educational Psychology, 1954, 45, 81-90.

Stanley, J. C. and Hopkins, K. D. Educational and psychological measurement and evaluation. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Swafford, J. O. Sex differences in first-year algebra. Journal for Research in Mathematics Education, 1980, 11, 335-346.

Thorndike, R. L. The problem of guessing. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971, 59-61.

Thorndike, R. L., Hagen E., and Wright, E. N. Canadian Cognitive Abilities Test, Form 1, Levels A-F. Toronto: Thomas Nelson and Sons, 1974.

Traub, R. E. and Fisher, C. W. On the equivalence of constructed-response and multiple-choice tests. _Applied Psychological Measurement_, 1977, _1_, 355-369.

Tulving, E. and Watkins, M. J. Continuity between recall and recognition. _American Journal of Psychology_, 1973, _86_ (4), 739-748.

Wolleat, P. L., Pedro, J. D., Becker, A. D., and Fennema, E. Sex differences in high school students' causal attributions of performance in mathematics. _Journal for Research in Mathematics Education_, 1980, _11_, 356-366.

APPENDIX A.

Copies of the Test Instruments and Table of Question Distribution

Table A.1

Distribution of Questions by Domain and Difficulty

Domain	Difficulty	Question Number						
Computation	High	8	9	14	26	27	33	38
	Low	1	2	7	12	15	18	29
Application	High	3	6	13	24	25	41	42
	Low	4	5	10	17	28	30	36
Algebra	High	16	20	31	32	34	37	40
	Low	11	19	21	22	23	35	39

NAME ______________________________

TEACHER'S NAME ______________________________

1. Divide: $45\overline{)1232}$

A) 25 remainder 7
B) 27 remainder 17
C) 29 remainder 27
D) 207 remainder 17
E) I don't know

2. Subtract: 51.2 - 4.35 =

A) 46.95
B) 46.85
C) 17.7
D) 7.7
E) I don't know

3. One fourth of a cake is shared equally among 3 children. What fraction of the whole cake did each of the children receive ?

A) $\frac{1}{7}$
B) $\frac{3}{4}$
C) $\frac{1}{12}$
D) $\frac{1}{3}$
E) I don't know

4. Seven pies are to be cut into fourths. How many pieces will there be ?

A) 14
B) 7
C) 28
D) 36
E) I don't know

5. The chart shows how long it took Ted to deliver papers last week. He worked a total of 320 minutes during the week. How long did it take him to deliver papers on Wednesday ?

Day	Mon	Tues	Wed	Thurs	Fri	Sat
Minutes	50	60	?	60	55	45

A) 54
B) 50
C) 55
D) 60
E) I don't know

6. A bicycle bought for $80.00 was sold at a loss of 30%. What was the selling price ?

A) $ 24
B) $104
C) $ 56
D) $ 30
E) I don't know

7. Multiply: $6 \times \frac{2}{3}$

A) 22
B) $20\frac{2}{3}$
C) 4
D) $6\frac{2}{3}$
E) I don't know

8. Divide: $\frac{3}{4} \div \frac{15}{8} =$

A) $1\frac{13}{32}$
B) $2\frac{1}{2}$
C) $\frac{2}{5}$
D) $1\frac{1}{2}$
E) I don't know

9. 12 is 15% of what number ?

A) 80
B) 180
C) 800
D) 1.8
E) I don't know

10. A map of B.C. is to be drawn so that 1 millimetre represents 5 kilometres. If the actual distance between Vernon and Penticton is 125 kilometres, how many millimetres apart should these two points be on the map ?

A) 125
B) 625
C) 120
D) 25
E) I don't know

11. Solve for n : $\frac{n}{4} = 8$

A) 32

B) 8

C) 2

D) $\frac{1}{32}$

E) I don't know

12. Calculate: $4^3 =$

A) 36

B) 64

C) 12

D) 32

E) I don't know

13. Patti took 20 pictures with her new camera. Five of the pictures were over-exposed and could not be developed. It cost $4.50 to develop the roll. What was the cost of each developed picture ?

A) 30¢

B) 18¢

C) $22\frac{1}{2}$¢

D) 25¢

E) I don't know

14. Divide: $0.0228 \div 0.003$

A) 7.6

B) 76.0

C) 13.0

D) 0.13

E) I don't know

15. Subtract: $12\frac{5}{6} - 3\frac{2}{3} =$

A) 9

B) $9\frac{1}{2}$

C) $16\frac{1}{2}$

D) $9\frac{1}{6}$

E) I don't know

16. In the formula $\frac{I}{PT} = R$ if I = 250, P = 1000 and T = 2 then R = ?

A) $\frac{1}{2}$

B) 50

C) 1

D) $\frac{1}{8}$

E) I don't know

17. If 37% of the Canadian population is under 20 years of age, what percent of the population is 20 years of age or older ?

A) 37%

B) 63%

C) 67%

D) 137%

E) I don't know

18. 0.95 as a percent is

A) 9.5%

B) 0.95 %

C) 95%

D) $9\frac{1}{2}$%

E) I don't know

19. Solve: 3x - 3 = 12

A) x = 7

B) x = 4

C) x = 3

D) x = 5

E) I don't know

20. Write an equation which represents the sentence:
" If 9 is added to 4 times a number the result is 29 ".

A) 4x = 29 + 9

B) 4(x + 9) = 29

C) 9x + 4 = 29

D) 4x + 9 = 29

E) I don't know

21. Write an expression which represents a number increased by 5.

A) 5 - x
B) x + 5
C) 5 > x
D) $\frac{5}{x}$
E) I don't know

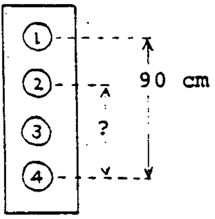
22. If n = 5 , then 2n + 4 =

A) 14
B) 18
C) 20
D) 11
E) I don't know

23. One number is 3 times as large as a second number. The sum of the two numbers is 72. What are the numbers ?

A) 24 and 8
B) 18 and 6
C) 12 and 36
D) 18 and 54
E) I don't know

24. A traffic signal has four equally spaced lights. How far apart are the centres of lights 2 and 4 ?



A) 22.5 cm
B) 30 cm
C) 45 cm
D) 60 cm
E) I don't know

25. A pasture is 48 m long and 30 m wide. How wide should a scale model of the pasture be if the length of the model is 24 cm ?

A) 15 cm
B) 38.4 cm
C) 60 cm
D) 12 cm
E) I don't know

26. Some of the digits have been covered. What digit was under the circle ?

```
  ▨ ▨ ◍ 2
- 3 4 8 5
---------
  ▨ ▨ 6 ▨
```

A) 1
B) 3
C) 5
D) 4
E) I don't know

27. Written as a decimal, $\frac{1}{8}$ =

A) 0.12
B) 0.8
C) 0.125
D) 0.18
E) I don't know

28. If a man mowed $\frac{2}{5}$ of his lawn, what part of his lawn does he still have to mow ?

A) $\frac{2}{5}$
B) $\frac{1}{5}$
C) $\frac{3}{5}$
D) 0
E) I don't know

29. Written as a decimal, four and four hundredths is :

A) 0.44
B) 44.00
C) 4.4
D) 4.04
E) I don't know

30. British Columbia became a province of Canada in 1871. Alberta became a province in 1905. How many years after British Columbia did Alberta become a province ?

A) 24
B) 134
C) 74
D) 34
E) I don't know

31. If 12(n + 7) = 108 then the value of n is

A) 9
B) 89
C) 2
D) $18\frac{5}{12}$
E) I don't know

32. If n is an odd number then the next odd number is:

A) n + 1
B) n + 2
C) n + 3
D) 2n - 1
E) I don't know

33. Which of these numbers is largest ?

$\left\{ \frac{2}{3}, \frac{4}{5}, \frac{3}{4}, \frac{5}{8} \right\}$

A) $\frac{2}{3}$
B) $\frac{4}{5}$
C) $\frac{3}{4}$
D) $\frac{5}{8}$
E) I don't know

34. If m = 2 and n = 3, then what is the value of 5(3m + 4n) ?

A) 35
B) 90
C) 85
D) 17
E) I don't know

35. What is the solution to 3n = 15 ?

A) 45
B) 18
C) 5
D) 12
E) I don't know

36. If one kg of oranges costs \$0.85, what will be the cost of 4.2 kg ?

A) \$4.55
B) \$4.85
C) \$3.98
D) \$3.57
E) I don't know

37. If $3n = 1$, then $n =$

A) 1
B) -2
C) $\frac{1}{3}$
D) 2
E) I don't know

38. Simplify: $\frac{0}{6} =$

A) 0
B) Infinity
C) 6
D) Cannot be done
E) I don't know

39. What is the solution of $2n + 8 = 20$?

A) 12
B) 14
C) 6
D) 10
E) I don't know

40. What values of n make the sentence $(n + 5) - 5 = n$ TRUE ?

A) 0 only
B) 0 and 5 only
C) all values of n
D) no values of n
E) I don't know

41. A salesman sold $2200.00 worth of merchandise in one month. If he earns 8% commission on sales, what is his commission for this month ?

A) $220.00
B) $176.00
C) $ 22.00
D) $ 17.60
E) I don't know

42. Paul earned $12 272 in twenty-six weeks. What was his weekly income ?

A) $482
B) $472
C) $468
D) $293
E) I don't know

NAME ______________________________

TEACHER'S NAME ______________________________

□□□□□□□□□

1. Divide: $45\overline{)1232}$ ____________________

2. Subtract: 51.2 - 4.35 = ____________________

3. One fourth of a cake is shared equally among 3 children. What fraction of the whole cake did each of the children receive ? ____________________

4. Seven pies are to be cut into fourths. How many pieces will there be ? ____________________

5. The chart shows how long it took Ted to deliver papers last week. He worked a total of 320 minutes during the week. How long did it take him to deliver papers on Wednesday ?

Day	Mon	Tues	Wed	Thurs	Fri	Sat
Minutes	50	60	?	60	55	45

____

6. A bicycle bought for $80.00 was sold at a loss of 30%. What was the selling price ?

____

7. Multiply: $6 \times \frac{2}{3}$

____

8. Divide: $\frac{3}{4} \div \frac{15}{8} =$

____

9. 12 is 15% of what number ?

____

10. A map of B.C. is to be drawn so that 1 millimetre represents 5 kilometres. If the actual distance between Vernon and Penticton is 125 kilometres, how many millimetres apart should these two points be on the map ?

____

11. Solve for n : $\frac{n}{4} = 8$

____

12. Calculate: $4^3 =$

____

13. Patti took 20 pictures with her new camera. Five of the pictures were over-exposed and could not be developed. It cost $4.50 to develop the roll. What was the cost of each developed picture ?

____

14. Divide: $0.0228 \div 0.003$

____

15. Subtract: $12\frac{5}{6} - 3\frac{2}{3} =$

____

16. In the formula $\frac{I}{PT} = R$ if $I = 250$, $P = 1000$ and $T = 2$ then $R = ?$

____

17. If 37% of the Canadian population is under 20 years of age, what percent of the population is 20 years of age or older ?

____

18. 0.95 as a percent is

____

19. Solve: $3x - 3 = 12$

____

20. Write an equation which represents the sentence:
" If 9 is added to 4 times a number the result is 29 ".

____

21. Write an expression which represents a number increased by 5.

____

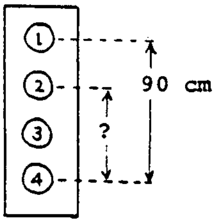
22. If n = 5 , then 2n + 4 =

____

23. One number is 3 times as large as a second number. The sum of the two numbers is 72. What are the numbers ?

____

24. A traffic signal has four equally spaced lights. How far apart are the centres of lights 2 and 4 ?



____

25. A pasture is 48 m long and 30 m wide. How wide should a scale model of the pasture be if the length of the model is 24 cm ?

____

26. Some of the digits have been covered. What digit was under the circle ?

```
  ▨▨●2
- 3485
------
  ▨▨6▨
```

____

27. Written as a decimal, $\frac{1}{8}$ =

____

28. If a man mowed $\frac{2}{5}$ of his lawn, what part of his lawn does he still have to mow ?

____

29. Written as a decimal, four and four hundredths is :

____

30. British Columbia became a province of Canada in 1871. Alberta became a province in 1905. How many years after British Columbia did Alberta become a province ?

____

31. If 12(n + 7) = 108 then the value of n is

____

32. If n is an odd number then the next odd number is:

____

33. Which of these numbers is largest ?

$\left\{ \frac{2}{3}, \frac{4}{5}, \frac{3}{4}, \frac{5}{8} \right\}$

____

34. If m = 2 and n = 3, then what is the value of 5(3m + 4n) ?

____

35. What is the solution to 3n = 15 ?

____

36. If one kg of oranges costs \$0.85, what will be the cost of 4.2 kg ?

____

37. If 3n = 1, then n =

____

38. Simplify: $\frac{0}{6} =$

____

39. What is the solution of $2n + 8 = 20$?

____

40. What values of n make the sentence $(n + 5) - 5 = n$ TRUE ?

____

41. A salesman sold $2200.00 worth of merchandise in one month. If he earns 8% commission on sales, what is his commission for this month ?

____

42. Paul earned $12 272 in twenty-six weeks. What was his weekly income ?

____

APPENDIX B.

Instructions to Test Administrators (First Testing Occasion)

First, let me thank you for taking time from your busy schedule to administer these tests to your Grade 7 students. The purpose of the study of which this testing is a part is to determine what effect, if any, that item format has on the score obtained by students on Mathematics achievement tests and how those scores are affected by gender, ability, item difficulty, and content. You will notice that the two tests are identical except that one test is in multiple-choice format and the other is in open-ended format.

The test is actually made up of six subtests of seven items each. Each of three content domains, Computation, Application, and Algebra, contain both easy and hard items mixed throughout the length of the test.

The study is a counterbalanced repeated measures design. Each student will take the test under both formats, one on April 29th in Math class, and the other two weeks later. On each occasion half the students will write one format and the other half will write the other format. Because of this, the tests must be identified with the students and also so that gender and ability can be coded along with the test results. Because the tests are given on two different occasions, it is important that you not alter your teaching plans based on the test content in the two-week period between the test dates. Please carry on as though the testing had not taken place. In addition, please make sure that you do not inform the students that they will be writing the same test again in two weeks time.

With this letter, you should receive enough copies of the test for your Grade 7 class. You will notice that the test formats are mixed. Please distribute the tests randomly to your class.

Once the tests have been distributed, please say to the class:

"Today you are going to write a test so that you can find out how well you write Math tests. Although this mark may not count as part of your total grade, I expect you to do your best. If you try hard on this test it is to your advantage. The results of these tests will be used to help teachers design better and fairer tests.

"Please write your full name, both your first name and your last name, on the front page of the test. Put my name (insert teacher's name) on the test as well.

"Now turn over the front page of the test. Each of you has a test which has 42 questions. Some of you have multiple-choice tests and some of you have tests where you must fill in a blank with the right answer. The questions on both tests are the same.

"To answer the multiple-choice questions, circle the letter of the best answer. If you have a fill in the blanks test then write your answer neatly on the line next to the question. You may do your working on the test, just make sure that your answer is neatly written in the proper place.

"You have one hour to finish the test. Do your best. Don't spend too much time on one question. You can always come back and answer it after you have finished the others. Check your paper before you finish.

When you finish please put your paper face down on your desk and sit quietly and read."

After the hour is up, please collect the papers and check that each student has put his or her full name on the test. I'll collect the tests on the test date or the day after. Please return all the tests used or unused.

If any students ask for help while writing the test, please do no more than read the question to them.

Instructions to Test Administrators

(Second Testing Occasion)

Thank you once again for agreeing to administer these tests to your Grade 7 students. Enclosed you will find a test for each student who took part in the first testing period. In addition, there are some blank tests for those who do not have complete tests or who were absent for the first testing period.

This second set of tests is to be administered on the day two weeks following the first testing period. On that day, please distribute the tests to the students. Any students who did not write the first test should be given one of the blank tests.

Once the tests have been distributed, please say to the class:

"Today you are going to write another test so that you can find out how well you write Math tests. Although this mark may not count as part of your total grade, I expect you to do your best. If you try hard on this test it is to your advantage. The results of these tests will be used to help teachers design better and fairer tests.

"Please make sure that you have the test with your name on it. If you received a blank test then write your full name and my name (teacher's name) on the front page of the test.

"Now turn over the front page of the test. Each of you has a test which has 42 questions. Some of you have multiple-choice tests and some of you have tests where you must fill in the blank with the right answer. The questions on both tests are the same.

"To answer the multiple-choice questions, circle the letter of the best answer. If you have a fill in the blanks test then write your answer neatly on the line next to the question. You may do your working on the test, just make sure that your answer is neatly written in the proper place.

"You have one hour to finish the test. Do your best. Don't spend too much time on one question. You can always come back and answer it after you have finished the others. Check your paper before you finish.

"When you finish please put your paper face down and sit quietly and read."

If some student's test needs to be replaced because of missing pages etc. please be sure to give the student a replacement of the same form. That is, if a student has a multiple-choice test the replace it with a multiple-choice test and if a student has an open-ended test then replace it with an open-ended test.

After the hour is up, please collect the papers. I'll collect the tests on the test date or the day after. Please return all tests, used or unused.

If any students ask for help while writing the test, please do no more than read the question to him.

APPENDIX C.

Summary Analysis of Variance Tables for Order of Administration, Class, and Item Difficulty

Table C.1

Summary Analysis of Variance
Class, Order, and Item Difficulty
Multiple-Choice Computation

Source of variance	Degrees of Freedom	Mean square	F
Mean	1	8591.477	2614.01*
Class (C)	8	12.990	3.95*
Order (O)	1	1.723	0.52
C x O	8	5.025	1.53
Error	195	3.287	
Difficulty	1	346.277	268.84*
D x C	8	2.919	2.27*
D x O	1	0.398	0.31
D x C x O	8	1.170	0.91
Error	195	1.288	

*p < .05

Table C.2

Summary Analysis of Variance

Class, Order, and Item Difficulty

Multiple-Choice Application

Source of variance	Degrees of Freedom	Mean square	F
Mean	1	9434.786	2484.68*
Class (C)	8	3.546	0.93
Order (O)	1	0.087	0.02
C x O	8	2.838	0.75
Error	195	3.797	
Difficulty	1	317.591	230.38*
D x C	8	1.009	0.73
D x O	1	2.646	1.92
D x C x O	8	0.656	0.48
Error	195	1.379	

*p < .05

Table C.3

Summary Analysis of Variance
Class, Order, and Item Difficulty
Multiple-Choice Algebra

Source of variance	Degrees of Freedom	Mean square	F
Mean	1	5843.306	1134.73*
Class (C)	8	13.437	2.61*
Order (O)	1	2.264	0.44
C x O	8	6.777	1.32
Error	195	5.150	
Difficulty	1	516.146	357.67*
D x C	8	1.382	0.96
D x O	1	0.991	0.69
D x C x O	8	1.061	0.74
Error	195	1.443	

*p < .05

Table C.4

Summary Analysis of Variance

Class, Order, and Item Difficulty

Constructed-Response Computation

Source of variance	Degrees of Freedom	Mean square	F
Mean	1	6620.177	2194.50*
Class (C)	8	19.306	6.40*
Order (O)	1	6.754	2.24
C x O	8	4.670	1.55
Error	195	3.017	
Difficulty	1	299.478	221.52*
D x C	8	2.779	2.06*
D x O	1	0.870	0.64
D x C x O	8	2.702	2.00*
Error	195	1.352	

*p < .05

Table C.5

Summary Analysis of Variance

Class, Order, and Item Difficulty

Constructed-Response Application

Source of variance	Degrees of Freedom	Mean square	F
Mean	1	6487.121	1429.61*
Class (C)	8	3.143	0.69
Order (O)	1	40.168	8.85*
C x O	8	3.992	0.88
Error	195	4.538	
Difficulty	1	499.965	449.90*
D x C	8	2.602	2.34*
D x O	1	1.347	1.21
D x C x O	8	0.739	0.67
Error	195	1.111	

*p < .05

Table C.6

Summary Analysis of Variance

Class, Order, and Item Difficulty

Constructed-Response Algebra

Source of variance	Degrees of Freedom	Mean square	F
Mean	1	2727.003	567.16*
Class (C)	8	9.468	1.97
Order (O)	1	72.963	15.17*
C x O	8	9.042	1.88
Error	195	4.808	
Difficulty	1	382.270	359.67*
D x C	8	1.926	1.81
D x O	1	2.295	2.16
D x C x O	8	0.613	0.58
Error	195	1.063	

*p < .05

APPENDIX D.

Table D.1

CELL MEANS AND STANDARD DEVIATIONS
COMPUTATION

			ABILITY	
FORMAT	DIFFICULTY	LOW	AVERAGE	HIGH
MALES		n = 31	n = 24	n = 33
M-C	HIGH	-1.39	-0.31	0.15
		(1.03)	(1.34)	(1.06)
	LOW	0.24	1.32	1.47
		(1.35)	(0.74)	(0.76)
C-R	HIGH	-1.90	-1.17	-0.36
		(1.03)	(1.19)	(1.12)
	LOW	-0.21	0.82	1.02
		(1.24)	(1.02)	(0.83)
FEMALES		n = 31	n = 40	n = 32
M-C	HIGH	-1.54	-0.59	0.50
		(1.07)	(1.40)	(1.05)
	LOW	0.95	0.98	1.77
		(0.91)	(1.23)	(0.79)
C-R	HIGH	-1.76	-0.80	-0.20
		(1.11)	(1.21)	(1.31)
	LOW	0.16	0.51	1.26
		(1.23)	(1.05)	(0.79)

Table D.2

CELL MEANS AND STANDARD DEVIATIONS
APPLICATION

			ABILITY	
FORMAT	DIFFICULTY	LOW	AVERAGE	HIGH
MALES		n = 31	n = 24	n = 33
M-C	HIGH	-1.05	-0.47	0.33
		(0.97)	(1.41)	(1.09)
	LOW	0.61	1.20	1.80
		(1.12)	(0.93)	(0.53)
C-R	HIGH	-1.77	-0.91	-0.51
		(1.03)	(1.33)	(1.24)
	LOW	0.20	1.21	1.23
		(1.07)	(0.71)	(0.74)
FEMALES		n = 31	n = 40	n = 32
M-C	HIGH	-1.06	-0.45	0.47
		(1.13)	(1.23)	(1.04)
	LOW	0.03	1.11	1.57
		(1.35)	(0.94)	(0.59)
C-R	HIGH	-2.17	-1.38	-0.39
		(1.05)	(1.26)	(1.36)
	LOW	-0.84	0.82	1.17
		(1.28)	(0.85)	(0.92)

Table D.3

CELL MEANS AND STANDARD DEVIATIONS
ALGEBRA

		ABILITY		
FORMAT	DIFFICULTY	LOW	AVERAGE	HIGH
MALES		n = 31	n = 24	n = 33
M-C	HIGH	-1.13	-0.42	0.52
		(0.82)	(1.24)	(1.07)
	LOW	0.66	1.30	1.59
		(1.20)	(1.01)	(0.84)
C-R	HIGH	-1.70	-1.05	-0.41
		(0.44)	(1.16)	(1.12)
	LOW	-0.67	0.66	1.06
		(1.02)	(0.94)	(0.96)
FEMALES		n = 31	n = 40	n = 32
M-C	HIGH	-1.20	-0.26	0.58
		(0.75)	(1.19)	(1.22)
	LOW	0.60	1.41	1.90
		(1.17)	(1.13)	(0.69)
C-R	HIGH	-1.59	-1.21	-0.34
		(0.67)	(0.94)	(0.99)
	LOW	-0.57	0.58	1.08
		(1.16)	(1.14)	(0.92)

Table D.4

CELL MEANS AND STANDARD DEVIATIONS
COMPUTATION
SCORES CORRECTED FOR GUESSING

			ABILITY	
FORMAT	DIFFICULTY	LOW	AVERAGE	HIGH
MALES		n = 31	n = 24	n = 33
M-C	HIGH	-1.76	-0.59	-0.12
		(1.09)	(1.52)	(1.19)
	LOW	0.05	1.17	1.40
		(1.44)	(0.89)	(0.83)
C-R	HIGH	-1.53	-0.88	-0.13
		(0.95)	(1.09)	(1.02)
	LOW	0.01	0.95	1.14
		(1.14)	(0.93)	(0.77)
FEMALES		n = 31	n = 40	n = 32
M-C	HIGH	-1.98	-0.92	0.31
		(1.21)	(1.52)	(1.16)
	LOW	0.84	0.83	1.73
		(0.99)	(1.35)	(0.85)
C-R	HIGH	-1.42	-0.54	0.01
		(1.01)	(1.11)	(1.20)
	LOW	0.34	0.65	1.35
		(1.13)	(0.96)	(0.73)

Table D.5

CELL MEANS AND STANDARD DEVIATIONS
APPLICATION
SCORES CORRECTED FOR GUESSING

			ABILITY	
FORMAT	DIFFICULTY	LOW	AVERAGE	HIGH
MALES		n = 31	n = 24	n = 33
M-C	HIGH	-1.35	-0.79	0.12
		(1.05)	(1.62)	(1.23)
	LOW	0.48	1.22	1.78
		(1.20)	(1.04)	(0.56)
C-R	HIGH	-1.48	-0.71	-0.31
		(0.95)	(1.23)	(1.13)
	LOW	0.33	1.26	1.30
		(0.99)	(0.66)	(0.69)
FEMALES		n = 31	n = 40	n = 32
M-C	HIGH	-1.43	-0.68	0.31
		(1.28)	(1.29)	(1.17)
	LOW	-0.17	1.02	1.55
		(1.47)	(1.03)	(0.64)
C-R	HIGH	-1.86	-1.12	-0.21
		(0.97)	(1.16)	(1.25)
	LOW	-0.63	0.90	1.23
		(1.18)	(0.79)	(0.86)

Table D.6

CELL MEANS AND STANDARD DEVIATIONS
ALGEBRA
SCORES CORRECTED FOR GUESSING

			ABILITY	
FORMAT	DIFFICULTY	LOW	AVERAGE	HIGH
MALES		n = 31	n = 24	n = 33
M-C	HIGH	-1.55	-0.72	0.31
		(0.87)	(1.37)	(1.18)
	LOW	0.50	1.17	1.50
		(1.25)	(1.06)	(0.92)
C-R	HIGH	-1.36	-0.78	-0.21
		(0.41)	(1.07)	(1.03)
	LOW	-0.40	0.80	1.15
		(0.94)	(0.88)	(0.90)
FEMALES		n = 31	n = 40	n = 32
M-C	HIGH	-1.61	-0.55	0.38
		(0.85)	(1.31)	(1.36)
	LOW	0.35	1.28	1.80
		(1.28)	(1.22)	(0.77)
C-R	HIGH	-1.28	-0.91	-0.13
		(0.62)	(0.86)	(0.90)
	LOW	-0.34	0.73	1.18
		(1.08)	(1.06)	(0.86)