COMPARISON OF DIFFERENT ITEM TYPES IN TERMS OF LATENT TRAIT IN MATHEMATICS ASSESSMENT

by

Zhen Wang

B.A., Shanghai University of Science and Technology, 1988
M.A., University of British Columbia, 1995

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES
(Department of Counselling, Educational Psychology and Special Education)

We accept this dissertation as confirming to the required standard

Dr. Nand Kishor..

Dr. Kadriye Ercikan..

Dr. Marshall Arlin..

THE UNIVERSITY OF BRITISH COLUMBIA

July 2002

Educational Psychology

Department of Counselling Psychology & Special Education

The University of British Columbia
Vancouver, Canada

Date July 21, 2002

COMPARISON OF DIFFERENT ITEM TYPES IN TERMS OF LATENT TRAIT IN MATHEMATICS ASSESSMENT

ABSTRACT

The question of whether multiple-choice (MC) and constructed-response (CR) items measure the same construct has not been satisfactorily addressed after many years of investigation. Previous studies comparing different formats in the domain of writing, reading, mathematics, and science have provided inconsistent results.

The purpose of the present study was to find out if format differences lead to multidimensionality in mathematics assessments. The study also added further to our knowledge about dimensionality. In addition, the study examined whether different formats assessed similar latent constructs based on Bloom's learning taxonomy. Because it is possible that the format differences may occur in one ability group and not in another ability group of examinees, the effects of format differences on the performances of students with different ability levels were also examined.

The four analyses reported in this investigation focused on mathematics assessments across different grade levels and time points of the school year. The four data sets were: (1) The Third International Mathematics and Science Survey (Grade 3 and 4, 1995); (2) The Third International Mathematics and Science Survey (Grade 7 and 8, 1995); (3) British Columbia Grade 12 Provincial Mathematics Examination (April 1998); and (4) British Columbia Grade 12 Provincial Mathematics Examination (August 1998).

Two different psychometric models were applied to address the research questions. First, Full-Information Item Factor Analysis (FIFA) (Bock & Aitkin, 1981), a combination of factor analysis and item response model analysis, was applied as an exploratory approach to detect the dimensionality of the test structures comprising different formats. Second, the Multidimensional Random Coefficient Multinomial Logit Model (MRCMLM) (Adams, Wilson, & Wang, 1997), a type of multidimensional latent trait model, was used as a confirmatory approach. In addition, analyses of cognitive demand and content were used to assist with the interpretation of the factor analyses. Two IRT based models (FIFA and MRCMLM), as well as examination of factor loadings, provided richer and stronger evidence than the single method applied in the previous studies on the investigation of format differences.

The findings indicated that one dominant trait was present in each data set. The existence of item local dependence and different cognitive demand (e.g., computation skill) were the most reasonable explanations for the existence of the non-significant minor dimensions. It appeared that the differences between MC and CR items did not affect the test structures. Most MC and CR items had high loadings on the same factor in the one-factor solution. In addition, MC and CR items correlated highly in the four sets of analyses. Therefore, the hypothesis that MC and CR items measured different mathematical proficiency was rejected. Additionally, MC and CR items did not differ in assessing students' cognitive ability beyond knowledge level. Such findings confirmed Hancock's (1996) conclusion that MC and CR items measured the same ability at each level of Bloom's cognitive framework.

High and low ability students differed in dealing with MC and CR item types. It appeared that the data set was two-dimensional for low ability students and unidimensional for high ability students in the three data sets (Data One, Data Three, Data Four). For the test (Data Two) that appeared to be more difficult than the above three tests, the data set was two-dimensional for high ability students and unidimensional for low ability students. In addition, low ability students performed better on MC items than on CR items, whereas high ability students performed similarly or better on CR items than on MC items in the three data sets.

It appeared that unidimensional IRT model can be used to calibrate those mathematics tests incorporating both MC and CR item types. However, the test structure may be two-dimensional (MC vs. CR) for the subgroups (high and low ability students). Therefore, reporting of two scores (MC vs. CR) sometimes may be useful for teachers and parents to diagnose students' weaknesses in the two formats. Teachers can thus find ways to help certain groups of students to improve their performances. The implications of the findings might be useful to teachers, test developers, and assessment specialists in making decisions in terms of item format selection, model selection, test development, scoring, and score reporting.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

APPENDIX

ACKNOWLEDGMENT

I would like to thank my dissertation supervisor Dr. Nand Kishor for his insight, guidance, and patience through each stage of the dissertation. I would also like to thank Dr. Kadriye Ercikan and Dr. Marshall Arlin for their guidance and comments. Discussions with them helped me to focus and clarify my thinking.

CHAPTER I: INTRODUCTION

The goal of educational testing is to find out how much students have learned or achieved so that various decisions can be made based on the results. Educators and assessment specialists have been studying and debating for a long time about which assessment format is a more appropriate measure in assessing students' academic ability. The reason for the concern is that test formats significantly influence education (Bennett & Ward, 1993). Test content and formats may provide cues to teachers about what is important to teach, and to students about what is important to learn (Lukhele, Thissen, & Wainer, 1994).

Among various types of assessment formats, the multiple choice (MC) item format is the most common in tests used for large-scale assessments, post-secondary admissions, job selection, college placement, certification, and many other assessment applications (Boodoo, 1993). According to Bennett, Ward, Rock, and LaHart (1990), the MC item format is narrowly defined as any item in which the examinee is required to choose an answer from a relatively small set of response options (e.g., four or five). The concept of constructed response (CR) item format is defined relatively broadly to include any item that requires the examinee to compose an answer.

MC items are often criticized for encouraging the teaching and learning of isolated facts and rote procedures at the expense of conceptual understanding and the development of problem-solving skills (Bennett & Ward, 1993). It is also argued that guessing may seriously contaminate the measurement. Proponents of CR items have argued that such items can measure traits that cannot be tapped by MC items, such as assessing dynamic cognitive processes (e.g., Fiske, 1990; Nickerson, 1989) and identifying students' misconceptions in diagnostic testing (Birenbaum & Tatsuoka, 1987). In addition, CR items may more closely represent real-world tasks so that students may be required to use many of the higher-order cognitive processes in order to formulate the answer (Bennett, 1993).

However, researchers have also argued that some MC items require complex skills from the respondent, whereas some CR items are relatively easy (Bennett & Ward, 1993). There are some limitations involved in using CR items. One such limitation is that only a small number of CR items can be administered in a typical testing period, thus reducing the breadth of content coverage. In addition, lack of standardization in test administration and objective scoring criteria may adversely affect comparability of results across examinees and tasks (Bennett & Ward, 1993). These conditions can threaten the representativeness of the test results as a sample of the individual's capabilities, and thus the validity and fairness of the test results.

Studies comparing different formats have been conducted in different domains such as: writing (Werts, Breland, Grandy, & Rock, 1980; Quellmalz, Capell, & Chou, 1982; Ackerman & Smith, 1988); verbal aptitude (Traub & Fisher, 1977; Ward, 1982; Van den Bergh, 1990; Janssen & De Boeck, 1996); and quantitative domain (Birenbaum & Tatsuoka, 1987; Bennett, Rock, Braun, Frye, Spohrer & Soloway, 1990; Bennett, Rock, & Wang, 1991; Bridgeman, 1992; Birenbaum, Tatsuoka, & Gutvirtz, 1992; Hancock, 1996; Wang & Wilson, 1996; O'Neil & Brown, 1998).

The conclusions from these studies comparing MC and CR formats in the different domains were equivocal. For example, three studies reviewed in the writing domain consistently revealed that there were two distinct factors related to MC and CR formats. However, by looking at the contradictory results from the eight studies in the quantitative domain, no consistent conclusion can be made regarding the format differences.

Traub and MacRury (1990) pointed out in their review paper that many studies in the last 70 years were seriously flawed in design and analysis. In some studies, the researchers used the correlation coefficient between the two item formats to investigate the format problem. As a result, these studies gave no direct information about the nature of the different traits assessed by the formats. Factor analytic studies that applied multiple measures of a particular trait or knowledge domain provided only limited evidence as to whether the two formats of the same content measured different characteristics (Traub, 1993).

Most studies between 1920 to 1970 employed methods such as direct correlation, the criterion correlation, and the treatment effect to investigate the relationship between MC and CR items. A criticism of these approaches was their lack of theoretical underpinnings (Hogan, 1981). In recent years, factor analysis, multitrait-multimethod confirmatory factor analysis, and item response modelling approaches have been applied more often to examine the format problem. The recent approaches were more theory-based and focused on latent "trait" comparisons under the assumption that it was the latent "traits" or "abilities" that determined how students responded to the test items. Besides, theory-based hypothesis testing also took "error variance" into account.

In the present study, two multidimensional IRT models, Full-Information Item Factor Analysis (FIFA) and Multidimensional Random Coefficient Multinomial Logit Model (MRCMLM), were applied as exploratory and confirmatory factor analyses to test whether MC and CR items were different. There are three advantages of applying the above two approaches: (1) they are multidimensional models; (2) statistical testing of the number of factors is possible using both approaches; (3) theoretical hypotheses can be tested and conclusions can be strengthened by using both exploratory and confirmatory approaches.

As Snow (1993) pointed out, there was no good demonstration to prove that CR items necessarily provided deeper, richer, more diagnostic assessment for instructional purposes than MC items; nor was there any evidence that MC items can be used to assess deeper and higher order thinking. Hence, this complicated issue needs further exploration.

From the literature review of the comparison of the format problem, it is not difficult to find that the nature of the format differences has not been fully explored. Hancock (1996) pointed out that the lack of precise identification of the cognitive skills was one of the serious problems in the previous studies. In order to clarify the nature of the differences between the formats, the analysis in the present study was based on the framework of Bloom's learning taxonomy (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956). Bloom's six levels of learning taxonomy were compressed into three levels: knowledge, complex procedures, and higher mental processes, which matched the table of specifications of the tests that were used in the present study. In addition, the nature of format differences might be clearer if we look at the cognitive demand involved. According to Snow and Lohman (1993), mathematical abilities were mainly represented by two factors, typically called *numerical facility* and *quantitative reasoning*. Mayer (1985), however, concluded in his review of the previous studies that students needed to have linguistic, factual, schema and strategic knowledge to solve mathematical problems. Thus, the examination of the cognitive demand of MC and CR items was done in the present investigation to help with the interpretation of the factor analysis results.

Format differences were investigated among different groups of students (high vs. low ability) as well. Snow and Lohman (1993) pointed out that learners' ability and experiences were important determiners of what attributes were measured by the test and how many different attributes were measured. Therefore, it is possible that the test structure can be unidimensional for one ability group and multidimensional for another ability group.

In the present study, it was hypothesized that high and low ability students differed in dealing with MC and CR items in the domain of mathematics. Some researchers believed that low ability students may apply "means-end strategy" and guess more often than high ability students do (Gagne, 1985). So, it is possible that high ability students may be able to handle both formats in a similar way because of their competence in the domain. However, other researchers (Ferguson, 1956; Anastasi, 1970) concluded that a test might be unidimensional for novices because all problems were relatively new for them and thus required some general problem-solving skills, whereas experts might show different patterns of skill development on different types of problems. It is obvious from these previous studies that high and low ability students differed in terms of the strategies they applied. However, the conclusions appeared to be inconsistent. Therefore, in the present investigation, whether high and low ability students handled different formats similarly was examined in order to explore further how assessment formats might impact different ability groups.

PURPOSE OF THE STUDY

The purpose of the study was threefold:

1. to determine whether differences between MC and CR item types may lead to some degree of multidimensionality;
2. to determine whether MC and CR items differ in the degree to which they assess similar cognitive levels based on Bloom's learning taxonomy;
3. to determine whether high and low ability students differ in dealing with MC and CR item types.

SIGNIFICANCE OF THE STUDY

Two psychometric methods were applied in the present investigation to assess whether the tests comprising both MC and CR items were unidimensional and whether MC and CR items measured two distinct constructs, namely MC and CR proficiency (Mislevy, 1993). Both of the two statistical models are multidimensional item response models. Full-Information Item Factor Analysis (FIFA) developed by Bock and Aitkin (1981) is a model that combines linear factor analysis and two- or three-parameter item response models. Multidimensional Random

Coefficient Multinomial Logit Model (MRCMLM) developed by Adams et al. (1997) is a Rasch-type item response model that can be used to test multidimensional models.

The advantages of using the multidimensional item response models over the studies that used other methodologies (e.g., linear factor analysis or unidimensional item response model) are threefold. First, they overcome the weakness of linear factor analysis. The linear factor analysis for dichotomous data based on tetrachoric correlation coefficients is not satisfactory because the matrix of sample tetrachoric correlation coefficients is almost never positive definite, and the coefficients become unstable as the values approach +1 or –1 (Bock, Gibbons, & Muraki, 1988). Second, the multidimensional models provide the possibility of exploring or confirming the dimensionality that is not possible by relying on the unidimensional item response model. Third, the applications of both exploratory (FIFA) and confirmatory (MRCMLM) approaches for the investigation of format differences may provide stronger evidence regarding the controversial format issues.

This study contributed at the level of theory, practice, and methodology. At the theoretical level, the differences between MC and CR items in terms of cognitive demand was investigated based on Bloom's learning taxonomy and Mayer's cognitive process model. So far, no such study has been found in the literature. At the practical level, the questions were answered as to whether a test was unidimensional and whether MC and CR items assessed different abilities. The hypothesis that high and low ability students differed in handling different formats was supported, which might have important implications for assessment specialists when they develop tests for different purposes (e.g., student selection or certification). At the methodological level, the study has revealed that the application of both exploratory FIFA and confirmatory MRCMLM provided richer evidence in terms of the dimensionality and format differences in mathematics assessment.

The generalizability of the results was enhanced by the following factors; first, the study was based on the data from three different grade groups: Grade 3 and Grade 4, Grade 7 and Grade 8, and Grade 12. Second, the study was based on samples from two large data banks: The Third International Mathematics and Science Survey (TIMSS, 1995) and the British Columbia Grade 12 Provincial Mathematics Examination (1998). Third, the test structures of the TIMSS and British Columbia Provincial Examination were different. Fourth, the tests were conducted at different time points: TIMSS was conducted in 1995 and the British Columbia Provincial Examinations were held in April and August of 1998.

The implications of the findings might be useful to teachers, test developers, and assessment specialists in making decisions in terms of item format selection, model selection, test development, calibration, and score reporting.

CHAPTER II: LITERATURE REVIEW

INTRODUCTION

The present review includes only recent studies from 1977 to 2001 when researchers started to follow more theoretical and sophisticated designs such as tau-equivalent measurements developed by Lord (1971) and Lord and Novick (1968). The studies selected for the review are searched through ERIC (Educational Research Index) using the keywords "multiple-choice" and "constructed-response." Another reason for the inclusion of these studies is that they were widely cited in the relevant literature.

The literature review is in four parts. First, the studies comparing different formats are reviewed within each domain (e.g., writing, language, quantitative domain, and science) because studies comparing different formats have been found in various domains, and the conclusions vary from domain to domain. Second, the methodological perspectives are also evaluated in comparison with the models applied in the study. Third, the literature regarding the theoretical framework of Bloom's learning taxonomy and cognitive theory is reviewed. Finally, the research studies on the differences between high and low ability groups in terms of cognitive ability and demand are examined.

WRITING

Studies that compared different formats in writing have been reported in recent years by Werts, Breland, Grandy, and Rock (1980); Quellmalz, Capell, and Chou (1982); and Ackerman and Smith (1988).

Werts et al. (1980) examined the relationships between direct (essay) and indirect (MC) measures of writing ability. Data were obtained on three different occasions for both types of measures from first-year undergraduate students (234 students with complete data). The indirect measure was the Test of Standard Written English (TSWE), a 30-minute test containing 50 MC items. The direct measures were 20-minute essays on three different topics.

A simplex model approach (Joreskog, 1974) was applied to test the hypotheses that the correlations among the factors would have a particular pattern and that the correlations among the essay errors might be equal. The analysis started with the hypothesis at each testing period. The three hypotheses were as follows: (1) the TSWE score and the essay ratings were measuring the same true score; (2) the correlation between the true scores at time one and time three was the product of the intervening correlation; and (3) the essay measurement errors were correlated. This model was found to be consistent with the data. The nonzero covariation (0.207) permitted in the model among the essay residual variables indicated that the essay formats measured some different characteristics (e.g., handwriting, spelling, and length of essay) in comparison to MC formats. However, it is not possible to obtain evidence from such a study as to what the nature of the format differences are. The nature of the format differences will be clear if cognitive processes involved are taken into account.

The study by Quellmalz et al. (1982) addressed two salient measurement issues concerned with the relationship between the structure of competency-based writing assessment tasks and the resulting performance. They tried to investigate (1) whether students' writing performance profiles were comparable on tasks differing in discourse mode (writing purpose), and (2) whether tasks requiring different response modes (paragraphs, essays, and MC items) provided the same type and quality of information about student writing competence.

Approximately 200 11th and 12th grade students from three high schools in a small school district in Los Angeles participated in the study. Students within each class were randomly assigned to one of four testing conditions defined by different discourse mode combinations for the essay tasks. All students were tested in three response modes: an MC test, a paragraph-length-writing sample (short CR), and the essay test (long CR).

The major comparisons reported were based on Multitrait-Multimethod (MTMM) confirmatory factor analysis models (Joreskog & Sorbom, 1978). The subscales of the scoring rubric formed the trait factors (General Impression, Focus, Organization, Scoring, Support, and Mechanics), and the discourse (narrative and expository) and response modes (multiple-choice, paragraph, and essay) formed the method factors. Hypotheses about trait and method influences on observed scores corresponding to specific (classes of) factor models were tested using analysis of covariance structures.

The results of the correlation, parametric, and MTMM analyses indicated that the levels of performance varied on tasks presenting different writing purposes. The MTMM model based on the two traits "Coherence (Focus and Organization scores) and Mechanics ratings" along with two factors "Narrative and Exposition" provided a good fit to the data ($\chi^2_{(17)} = 18.68$,

$p > 0.05$). The findings suggested that generalizations about student writing competence must reference the particular discourse domain rather than the general domain of writing. In addition, tests of response mode effects within the MTMM confirmatory factor-analytic framework showed method variance to be present in varying degrees. The final MTMM model based on three trait factors "Coherence, Support, and Mechanics" and four method factors "essay1, essay2, paragraph, and MC) fit the response mode data quite adequately ($\chi^2_{(83)} = 95.55$, $p > 0.05$). The MTMM framework was complex; however, it provided an appropriate approach in dealing with the issue of item formats.

In a more recent study done, Ackerman and Smith (1988) investigated the unique skills and abilities measured by different assessment formats in writing. Particularly, they based their study on the conceptual framework of writing process proposed by Hayes and Flower (1980). In the study, Hayes and Flower's model was modified as a framework for examining the component differences between the processes involved in direct and indirect assessment of writing. Ackerman and Smith hypothesized that the factor structure of the indirect method of assessment would differ from that of the direct method.

Two hundred and nineteen 10th grade students from a parochial high school in Southeastern Wisconsin participated. Students were randomly selected from traditional English classes.

Confirmatory factor analyses were done using LISREL IV. An eight-factor CFA was imposed on all 12-subtest scores for the MC and CR tests. The first six factors were specified as trait factors (Spelling, Capitalization/Punctuation, Correct Expression, Usage, Paragraph Development, and Paragraph Structure). The seventh factor was hypothesized to be a "recognition" factor. The eighth factor was targeted to be the "generation" factor, and only the six sub-scores of CR test were free to load on it. The results of the confirmatory factor analysis indicated a reasonably good fit: $\chi^2_{(24)} = 20.4$, $p > 0.05$. The recognition factor was dominated by MC subscores, and the generation factor was clearly dominated by the subscores (Usage) of CR items. Another nine-factor CFA model was tested and included essay items that were free to load on the organization factor. The results indicated that the model was a very good fit to the data: $\chi^2_{(84)} = 71.48$, $p > 0.05$. MC and CR loading dominated the recognition factor. The organization factor was clearly dominated by the Paragraph Development (0.71) and Paragraph Structure (0.42).

The authors concluded that scores obtained from direct and indirect methods of writing assessment provided different information. Generation skills especially can be more accurately assessed with an essay task.

A strength of the study lies in the fact that it incorporated the writing process framework (recognition, generation, and organization factors) so that the nature of the differences between assessment formats can be explained quite well. However, the study is limited in its conclusion by relying on only confirmatory factor analysis method.

Generally speaking, all of the three studies in the writing domain consistently showed two distinct characteristics related to MC and CR formats. However, the nature of the differences between MC and CR items has been investigated only by Ackerman and Smith (1988) who looked into the writing processes in which students might be involved when handling the different formats.

VERBAL APTITUDE

Four studies have been selected for review in the domain of verbal aptitude (sentence completion, antonyms, synonyms, and reading). They were done by Traub and Fisher (1977), Ward (1982), Van den Bergh (1990), and Janssen and De Boeck (1996).

Traub and Fisher (1977) investigated the equivalence of the three response formats (CR, standard MC, Coombs' MC) in two different content domains (mathematics and verbal aptitude). In addition, the study was designed to identify format factors that were defined as factors associated with the tests employing the same response format regardless of test content.

The examinees in the study were 199 8th grade students (93 females). Lord's procedure for testing the equivalence of measures and the method of confirmatory factor analysis were applied to the data obtained from a single group of subjects. Lord (1971) hypothesized that two sets of measurement would differ only because of errors of measurement, differing units of measurement, and differing arbitrary origins for measurement.

The main conclusion of this study concerns the equivalence of measurements arising from the tests employing different formats based on the same content. When content was held constant and allowance was made for differences due to errors of measurement and scale parameters, the tests of mathematical reasoning were equivalent regardless of format, but the tests of verbal comprehension were not. Another conclusion was that the existence of a format factor was not justified because the loadings were small in absolute magnitude (from 0.01 to 0.32).

The strength of the study lay in its control of the item content, and confirmatory factor analysis was also appropriate for the purpose of the study. However, a limitation of the study was that the results were based on the factor analysis alone.

Ward (1982) applied a factor-analytic examination of the influence of response format. Each of three item types was given in each of the four formats, varying in the degree to which they required production of answers. The fit of the data to each of the two ideal types of factors was examined: item-type factor and format factor.

Three item types were employed: antonym items, sentence completions, and analogies. Three formats in addition to the MC items were used: single-answer format, multiple-answer format, and keylist format. Factor analysis and MTMM analysis were applied to test the hypotheses. The results of the principal component analysis indicated that only one single factor represented the data. The first factor accounted for 57% of the total variance and the next largest accounted for only 7% of the variance. By looking at the loadings that ranged from 0.57 to 0.86 on the first factor of the principal axes factor analysis followed by the varimax rotation, Ward concluded that the completion and antonym items measured the same attribute regardless of the format in which the items were administered.

In MTMM analysis, each of the three item types was regarded as a “trait” and each of the four response formats constituted a "method." The correlation coefficients of the MTMM showed no evidence of the CR tests clustering according to the response format. Ward concluded that discrete verbal item types appeared to measure essentially the same abilities regardless of which format was used.

Van den Bergh (1990) explored the question of whether items for reading comprehension were congeneric regardless of the format of the items. Congenericity (Joreskog, 1974) means that the tests measure the same traits except for errors in measurement. The framework of the Structure-of-Intellect (SI) model (Guilford, 1971) was applied. In the SI model, abilities were defined on three dimensions: operations, content, and product. It was hypothesized that CR questions for reading comprehension measured divergent- and convergent-production abilities, whereas cognition and evaluation abilities were measured by MC items. These differences were observed in differences in the regression weights of reading comprehension scores on the SI ability scores. Another question investigated was whether SI abilities were involved in answering items in traditional reading comprehension tests. It was hypothesized that memory abilities were crucial for answering the reading comprehension items, but were not differentially involved in answering CR and MC items.

Five hundred ninety 3rd grade children from 12 different Dutch high schools were selected. The structural equation modelling approach (LISREL) was applied in the study. Three models were tested. In the first model, it was assumed that there were no differences in regression of the reading comprehension factor on the SI factors due to the reading comprehension test. It was also assumed that there were no differences between the tests due to the format of the items. In the second model, the format model, the restrictions were loosened a bit; differences in regression weights due to the reading comprehension tests were allowed, but again no differences between formats were permitted. In the third model, no restrictions were placed on the regression of the reading comprehension factor on the cognition, evaluation, convergent-production, and divergent-production ability factors. Differences in regression weights as well as differences in residual variances were allowed. In this "difference model," the regressions of the memory abilities were constrained over item format.

The results indicated that the fit index between Model 2 and Model 3 (GFI = 0.925, RMSE = 0.081 vs. GFI = 0.946, RMSE = 0.077) did not differ much due to the format of the items. Therefore, Van den Bergh (1990) concluded that CR and MC items for reading comprehension were evidently congeneric with respect to the SI abilities measured. Because a relatively large proportion of the variance in true reading comprehension scores was accounted for by the SI ability tests, it was tempting to conclude that semantic abilities involved in answering reading comprehension items with a different format did not differ much.

A strength of the study lay in the fact that a theoretical framework (intellectual abilities) was established while the item formats were examined. Therefore, the author was able to investigate how intellectual abilities were differentially involved in answering CR and MC items.

A validation study might be useful to confirm the results from this study. A recent study done by Janssen and De Boeck (1996) made an effort to incorporate a cognitive approach to the study of format differences using synonym tasks. Another purpose of the study was to contribute to the discussion about the effect of using the same item stems across formats (item families).

The synonym tasks were administered both with and without item families. The total item set available was partitioned into three tests of 40 items for the generation, evaluation, and free-response synonym tasks (CR items), respectively. The score on each task was equal to the number of correct items. All the data were used simultaneously to estimate a structural equation model for multiple regression with latent variables. The scores were derived from the generation, evaluation, and free-response synonym tasks (CR items). Each of them was defined as a separate factor. Maximum likelihood estimation procedure was used for the estimation.

Both the generation factor and the evaluation factor contributed significantly to the prediction of the free-response synonym factor (0.58 and 0.20, respectively). The results showed that a type of free-response synonym task (CR items) existed in which a response-production factor played an important role when evaluation was controlled. In sum, the findings indicated that a response-production component was involved in answering a free-response synonym task (CR items). At the methodological level, the results of the study suggested that a study of format differences should use a design that controls both for item-family effects and for content effects.

In sum, three of the four studies reviewed in the verbal domain consistently indicated that format effect did not exist. However, Janssen and De Boeck (1996) in their study suggested that there was a response-production factor in the free-response synonym task (CR items).

QUANTITATIVE DOMAIN

Eight studies were reviewed in the quantitative domain (computer programming, mathematics). They were carried out by Birenbaum and Tatsuoka (1987), Bennett, Rock, Braun, Frye, Spohrer, and Soloway (1990), Bennett, Rock, and Wang (1991), Bridgeman (1992), Birenbaum, Tatsuoka, and Gutvirtz (1992), Hancock (1996), Wang and Wilson (1996), and O'Neil and Brown (1998).

The study by Birenbaum and Tatsuoka (1987) examined the equivalence of the two formats for diagnostic purposes. The study evaluated the effect of the response format (MC vs. CR) on the rules of operation underlying examinees' response patterns in fraction-addition arithmetic items.

A test in fraction addition was administered to 285 eighth-grade students, 148 of who responded to the CR version of the test and 137 to the MC version. The two data sets were compared in terms of the underlying structures of the test, the number of different error types, and the diagnosed sources of misconception (bugs) reflected in response patterns. The results indicated considerable differences between the two formats, with more favourable results for the CR format. The underlying structure, which was examined by smallest space analysis, seemed clearer in the CR data set where the configuration of the items in the two-dimensional space clearly indicated two clusters: one cluster of items with like denominators and the other cluster of items with unlike denominators. The item configuration of the MC dataset, on the other hand, seemed quite diffuse, with no distinct separation between the different item types.

It seemed that the cognitive processes involved in these two response formats were quite different. The implications for diagnostic achievement testing in procedural tasks were obvious. CR tests may provide the appropriate information for identifying students' misconceptions with respect to the given subject matter. The MC format may not be appropriate for this purpose.

The purpose of the study by Bennett et al. (1990) was to assess the relationship between the expert system scored constrained CR item type and MC and CR items contained on the College Board's Advanced Placement Computer Science (APCS) Examination. The magnitude of this relationship was central to evaluating the potential of this item type as an eventual replacement for more CR formats and as a supplement to MC question.

Confirmatory factor analysis was used to test the fit of a three-factor model to these data. Three alternative methods were compared: (1) a null model in which no common factors were presumed to underlie the data; (2) a general model in which all variables loaded on a single factor; and (3) a two-factor solution in which APCS test was different from the constrained CR items.

The results of the hierarchical chi-square tests suggested that the three item types formed a single factor in one sample (one factor vs. null: χ^2 difference = 1514.48, $p < 0.01$). However, a two-factor model with faulty solutions defining a separate factor might better account for the data in the second sample (two- vs. one-factor: χ^2 difference = 12.36, $p < 0.01$). In addition, the factor correlations indicated that the expert system scored constrained CR item was highly related to both CR and MC items.

Another study done by Bennett et al. (1991) was intended to assess the equivalence of MC and CR items in computer science. Subjects were two samples of 1,000 students randomly drawn from the population of 7,372 high school students taking the Advanced Placement

Computer Science (APCS) examination. The instrument included both MC (50 items) and CR (5 items).

Confirmatory factor analysis was used to test the fit of a two-factor model where each item format marked its own factor. Parcels of MC items marked the first factor, whereas the second factor was indicated by the five CR problems. The results of the comparison between the two-factor model with the alternative hypothesis of one-factor model indicated a statistically significant improvement for the two-factor over the one-factor model in both samples (χ^2 difference = 10.46, $p < 0.01$ for sample 1, χ^2 difference = 40.57, $p < 0.01$ for sample 2). However, since the gain of two- over one-factor model was small by other fit indices (e.g., AIC), Bennett et al. concluded that the one-factor model provided the most parsimonious fit in both samples. Such evidence did not agree with the belief that MC and CR formats measured substantially different constructs (i.e., trivial factual recognition vs. higher-order processes).

The study by Bridgeman (1992) compared open-ended (CR items) and corresponding MC formats in the domain of quantitative items (Graduate Record Examination-Q). The main question addressed was the extent to which the CR versions of the items paralleled the MC versions in terms of difficulty, discrimination, and correlational structure.

A method based on item response theory was applied for the graphical comparison of the performance of items in the MC and grid-in formats. Three-parameter item characteristic curves were computed using the LOGIST program.

The results indicated that, at the level of the individual item, there were striking differences between the open-ended (CR) and MC formats. Some items that were relatively easy in the MC format were relatively difficult in the open-ended format (CR). Item characteristic curves for questions in the grid-in and MC formats were nearly overlapping for some items and highly discrepant for others. Format effects appeared to be particularly large when the MC options were not an accurate reflection of the errors actually made by students. In addition, it became clear that asking for a rounded answer in the grid-in format was not equivalent to asking for an approximation in MC format. However, total test scores in both formats appeared to be comparable. Both formats ranked the relative abilities of students in the same order. Gender and ethnic differences were neither lessened nor exaggerated. Correlation with other test scores and college grades were about the same.

Birenbaum et al. (1992) further examined the effect of response format using additional diagnostic assessment criteria. Comparisons were made between parallel-stem items with identical response format and stem-equivalent items with different response formats. In addition to comparing the diagnostic results of different formats, parallel CR items were contrasted to address the issue of "bug" instability. A bug is an incorrect rule that an examinee uses to solve a problem. The bugs are often unstable so that the diagnostic results of subsets with the same format or different formats may be affected.

The results of the bug and rule-space analyses yielded similar results with respect to the format effect. Both analyses indicated a closer similarity between the two parallel CR sub-tests than between the stem-equivalent CR and MC sub-tests. On the average, the MC sub sets tended to be more difficult and had lower internal consistency reliabilities than the two CR sub-tests.

The study by Hancock (1996) investigated the comparability with which the MC and CR formats assessed particular levels of complexity within the cognitive domain. Specifically, by using the framework provided by Bloom's taxonomy, the author examined the degree to which MC and CR tests measured the same cognitive skills.

Correlation method and factor analyses were applied to analyze the data. A pairwise comparison of the MC sub-tests with their corresponding CR measures at each taxonomic level showed fairly high corrected correlation at all taxonomic levels. Given sound test construction, MC items appeared to be able to measure the same abilities as CR items across the first four levels of Bloom's taxonomy.

Wang and Wilson (1996) compared MC items and CR (performance-based) items in mathematics using the Random Coefficients Multinomial Logit Model (Adams & Wilson, 1996). The data were selected from the California Learning Assessment System. Each form contained 20 MC items, two open-ended (CR) items, and one investigation (CR) item. The MC items and CR (performance-based) items were compared in terms of the information function and standard errors. The results indicated that a CR item provided about 4.5 times more information than an MC item on average. Thus the CR items were more accurate in detecting abler persons than the MC items.

O'Neil and Brown (1998) investigated the effect of item format on metacognitive and affective processes of children in the context of a large-scale mathematics assessment program. A sample of 1032 8th-grade students were selected to participate in a mathematics examination including both MC and CR formats as part of the California Learning Assessment System. Metacognition and affect questionnaires were then administered to the students following each format.

The results indicated that open-ended and MC question formats have differential effects. Open-ended questions induced more cognitive strategy usage, less self-checking, and greater worry than did MC questions. These effects did not vary substantially as a function of gender and ethnicity.

The strength of the study lay in the fact that the researchers incorporated metacognition (planning, self-checking, awareness, and cognitive strategy) when investigating the format differences. The reason that students performed better on MC than CR items seemed to be more clearly explained in this study than in other studies.

By looking at the contradicting results from the above eight studies in the quantitative domain, it is impossible to conclude whether MC and CR items measure the same thing. The equivocal results may, in part, be due to the different methodologies used.

SCIENCE

Fewer studies have been conducted to investigate the format differences in the domain of science: Harke, Herron, and Lefler (1972), Martinez (1991), Lukhele, Thissen, and Wainer, (1994); and Thissen, Wainer, and Wang (1994).

In the study by Harke et al. (1972), a repeated measure design was used to determine the similarity between the measurements of achievement on the randomized MC format and the conventional manually graded written answer format. The high correlation (.73) between the scores on the written solution and MC items in the physics problem test indicated that the MC format could be used as an adequate substitute for the universally accepted written solution tests in physics.

Martinez (1991) did a study comparing MC items with figural response items. He pointed out that figural response items differed from MC in that figural items required CR and depended on figural material. This study contrasted MC and CR (figural) items by using item and test statistics. Subjects from Grades 4, 8, and 12 were drawn from a national sample of students. Twenty-five CR (figural) items were written in three content areas: life sciences, physical sciences, and earth and space sciences.

In general, CR (figural) items were more difficult than MC items, but the differences in difficulty interacted with the relative difficulty of the items. The item and test statistics showed that CR (figural) items were generally comparable or superior to parallel MC items.

The purpose of the investigation by Lukhele et al. (1994) was to compare the relative value of MC and CR items. The data from College Board's Advanced Placement examinations in chemistry and United States history were analyzed based on fitting item response models. The three-parameter logistic IRT model was used for MC items, and Samejima's graded model (1972) and Bock's nominal model (1972) were used for CR items. The results of the comparison of the standard errors of estimate of proficiency of the two tests (75 MC vs. 75 MC + 2 CR items) showed that there was a marginal gain in precision obtained by including the two CR items. In terms of the comparison of the information function, the CR items provided less information in more time at greater cost than did the MC items.

Thissen et al. (1994) investigated whether measuring with CR items was the same thing as measuring with MC questions, and whether it was meaningful to combine the scores on the CR sections with the MC scores to yield a single reported total score. Restricted factor analysis was applied on the data collected in the domains of computer science and chemistry tests of the College Board's Advanced Placement program.

The two-factor model showed that for the most part the CR sections measured the same underlying proficiency as the MC sections. However, there was also a significant yet small amount of local dependence among the CR items that produced a small degree of multidimensionality for each test. If two items are locally dependent, then it means that the success on one item is dependent on the other item. CR items and performance assessments appeared likely to produce far more local dependence than what is produced by traditional MC tests.

Contradictory results were found in the science domain. It appeared that those different methods and different sub-domains were the main reasons for the different conclusions. Future research studies applying different methods are needed to compare formats across sub-domains in science.

A STUDY ACROSS DOMAINS

Ercikan, Schwarz, Julian, Burket, Weber, and Link (1998) examined whether MC and CR items designed to assess similar constructs can be calibrated together to create a common scale. Empirical results were provided, using data from different subject areas such as reading, language, mathematics, and science.

The researchers applied a three-parameter logistic (3PL) model (Lord, 1980) for the MC items and a two-parameter partial credit (2PPC) model (Yen, 1993) for the CR items. Eight hundred students from Grades 3, 5, and 8 participated in the reading, language, mathematics, and science tests that were developed. The test contained both MC and CR items.

The results from the study indicated that MC and CR items assessed similar latent constructs that were sufficiently similar to allow the creation of a common scale and provide a single set of scores for responses to both item types. Although simultaneous calibrations led to loss of information for CR items, the differences in information loss were negligible in most tests, and all the large differences were due to local dependence. Additionally, these researchers found that separate calibration of MC and CR item types lead to poor measurement accuracy due to the CR sections (e.g., extreme test difficulty, short test length). Therefore, the authors concluded that combining MC and CR item types increased overall measurement accuracy.

The study provided strong evidence regarding the comparability of MC and CR item types across the four different subjects and at three grade levels. However, because the results from the study were based on one IRT method, it would be valuable to compare calibration of the two item types using other IRT models.

LATENT TRAIT MODELS AND LEARNING TAXONOMY

Latent Trait Models

The two latent trait models that were tested in the present study were FIFA (Bock et al., 1988) and MRCMLM (Adams et al., 1997). Based on the literature review, it seemed that the most frequently used methodology in the past was linear factor analysis (Quellmalz et al., 1982; Ackerman & Smith, 1988; Bennett et al., 1990; Hancock, 1996). Linear factor analysis has been a powerful and indispensable tool for exploring underlying relationships among a set of continuous variables for many years. However, for dichotomous item responses, the tetrachoric correlation coefficients become unstable as they approach extreme values. The coefficient cannot be computed from a four-fold table with a zero frequency in any cell. Christoffersson (1975), Muthen (1978), and Bock and Aitkin (1981) provided the solutions to the problem with tetrachoric factor analysis. However, the FIFA, which is based on multidimensional item response models is regarded as a better approach for exploring dimensionality because the other two approaches become computationally heavier as the number of items increases (Muthen, 1984). FIFA uses all the information available in the matrix of dichotomously scored response patterns. FIFA is not limited by the number of items, and is applicable to exploring the dimensionality of long tests. Based on Muraki and Engelhard's (1985) simulation study comparing FIFA with a conventional tetrachoric factor analysis, FIFA did better in estimating factor loadings in both no-guessing and guessing data sets, especially when item difficulties varied to a great extent.

Item response models have also been applied to compare MC and CR items (Bridgeman, 1992; Birenbaum & Tatsuoka, 1987; Wang & Wilson, 1996; Ercikan et al., 1998). The one-, two-, or three-parameter item response models have their limitations when the test is multidimensional. The applications of the Rasch-based MRCMLM helped to explore the format problem in a confirmatory perspective. Applications of most multidimensional item response models are limited because they are very complex and the parameters are not easily interpretable, but MRCMLM offers some advantages because of its simplicity of interpretation. Wang (1994) found that MRCMLM worked better than separate unidimensional Rasch model analyses because subscales were estimated simultaneously and all the collateral information was taken into the estimation procedure.

Both FIFA and MRCMLM are IRT-based models that overcome the problems of factor analysis and unidimensional IRT models. Consequently, they were both applied in the investigation of format problem in the present study. FIFA was used as an exploratory approach, and MRCMLM was applied as a confirmatory approach.

Learning Taxonomy

From the review of literature on the comparison of the format differences, it is obvious that the nature of the format differences has not been fully explored. Hancock (1996) pointed out that the lack of precise identification of the cognitive skills was one of the serious problems in the previous studies. Therefore, it is difficult to determine whether CR items can tap higher levels of cognitive skills than MC items. In order to clarify the nature of the differences between the formats, it is important to establish a framework for defining such cognitive skills.

Bloom et al. (1956) established the learning taxonomy that provided a basis for building curriculum and tests. Bloom's six levels of taxonomy are *knowledge*, *comprehension*, *application*, *analysis*, *synthesis*, and *evaluation*. The simplest and basic level of the taxonomy — *knowledge* involves the recall of specifics and universals, the recall of methods and processes, or the recall of a pattern, structure, or setting. The higher levels of the taxonomy include intellectual abilities and skills that are hierarchical in nature, from *comprehension*, *application*, *analysis*, and *synthesis* to *evaluation*. *Comprehension* refers to a type of understanding in which the individual knows what is being communicated. *Application* refers to the use of abstractions in particular and concrete situations. *Analysis* involves the breakdown of a communication into its constituent elements or parts so that the relative hierarchy of ideas is made clear. *Synthesis* involves the process of putting together elements and parts in order to arrange and combine them so that they constitute a pattern or structure. *Evaluation* involves quantitative and qualitative judgment about the extent to which material and methods satisfy criteria. Generally speaking, Bloom and his associates postulated that the six levels showed an order from simple to complex and that the successive levels were cumulative in the higher levels, building upon and incorporating the lower. However, the linear assumption has been criticized by people who argued that the assumption was too simplistic (Ormell, 1979). Ormell pointed out that certain demands for knowledge were more complex than certain demands for analysis or evaluation. Bloom et al. (1956) acknowledged that it was not possible to make distinctions as clear-cut as one would like because inversions sometimes occurred and there was frequent overlap between and within categories.

One feature of Bloom's taxonomy was followed in the present study. Bloom made a special distinction between the lowest level, knowledge, and the five higher levels. He asserted that knowledge involved "little more than the remembering of the idea or phenomenon in a form very close to that in which it was originally encountered." In contrast, the higher levels of the taxonomy were presumed to involve the exercise of intellectual skills and abilities on that knowledge.

In order to find out if CR items differ from MC items in tapping students' ability, it is important to use a framework of such learning taxonomy so that the nature of the differences between formats can be compared at the same level of cognitive complexity.

Examination of Factor Loadings

In order to find out more about the nature of format differences, it is useful to incorporate examination of factor loadings and analyses of the cognitive demand that items may place on the students. According to Snow and Lohman (1993), cognitive analyses of existing measures can help to improve understanding of the constructs represented by those items. Sources of item difficulty might be understood and manipulated in new ways. The processing skills and knowledge components of test performance that do and do not produce individual differences in scores could be clarified. Therefore, it is useful to use cognitive analysis as a new source of evidence for construct validation. A test item may require both declarative and procedural knowledge. However, the demands may differ from one item to another depending on the context. For some people, a knowledgeable or skillful performance of a specific item requires conscious attention resources for successful operation. For others, automatization occurs.

Very few researchers have investigated format differences by looking at the cognitive demand or processes involved. Ackerman and Smith (1988) examined the component differences between the processes involved in direct and indirect assessment of writing based on the writing process framework (recognition, generation, and organization factors). As a result, the nature of the differences between assessment formats can be explained quite well. However, few similar format comparison studies have been found in the domain of mathematics. One recent study by O'Neil and Brown (1998) confirmed that open-ended questions induced more cognitive strategy usage, less self-checking, and greater worry than MC items in mathematics assessment. According to the studies on cognitive structures and processes in the mathematics domain, mathematical ability was defined as numerical facility and quantitative reasoning (Snow & Lohman, 1993). Numerical facility was defined as performing simple arithmetic computations. Quantitative reasoning included the ability to assemble a plan to solve a problem.

Mayer (1985) suggested that linguistic knowledge, and factual knowledge as well as schema knowledge would be needed for the problem representation (translation and integration). Factual and linguistic knowledge would be needed for problem translation, and schema knowledge would be needed for integration. Schema knowledge represents student's knowledge of the form of the problem (e.g., equation of the structure of the problem). Schema knowledge assists students in integrating propositions in several ways. First, schema knowledge can help to define slots for particular pieces of information and relations among these facts, which can greatly assist in correctly organizing propositions (Mayer, 1982). Second, schema knowledge can help to get rid of irrelevant information when a problem type is recognized. Therefore, in addition to linguistic and factual knowledge, the student needs schema knowledge in order to put the variables together in a coherent way. According to the previous studies, many of the difficulties students have in solving story problems can be due to the wrong schema (Loftus & Suppes, 1972; Hinsley, Hayes, & Simon, 1977).

To solve the problem, students need to know the rules (e.g., arithmetic or algebra) as well. Sometimes students need to use strategy knowledge (e.g., working backwards) to control and use the knowledge at the right time (Mayer, 1980). In the present investigation, the cognitive demand that an item might place on students was analyzed using Mayer's models (1985).

HIGH ABILITY AND LOW ABILITY STUDENTS

Many psychologists and educational researchers have been interested in finding out the differences between experts and novices in terms of knowledge organization and problem-solving strategies. In the mathematics domain, the knowledge organizations of experts and novices in mathematics differ (Gagne, 1985). Some researchers have concluded that experts' knowledge organization is more consistent with the accepted structure of subject matter than is novices' knowledge organization (Gagne, 1985; Geeslin & Shavelson, 1975; Silver, 1981; Feltovich, 1981). Feltovich (1981) found that an expert's structure seems to be more interconnected and more hierarchical than a novice's. Researchers believe that organization is very important because it influences what people attend to in the problems, what they recall about problems, and therefore how they solve the problems.

In terms of problem-solving strategies, researchers have found that novices, not experts, use what is considered to be the more powerful strategy (Chi, Glaser & Rees, 1982). Experts would use the simpler working-forward strategy and novices would apply the more powerful means-ends analysis. The expert's working-forward strategy is believed to be based on a great deal of prior learning experiences. In addition, the expert is automatically recognizing known patterns and applying the action sequences associated with those patterns (Gagne, 1985).

However, some other earlier researchers seem to have the opposite opinion. According to Ferguson (1956), novices might use the same general problem-solving skills because all the problems are relatively novel for them, whereas experts might show different patterns of skill development on different types of problems. Anastasi (1970) supported the above conclusions by pointing out that abilities tend to differentiate with practice.

Based on the previous studies, it seems that high and low ability students are very likely to differ in their knowledge structure and the strategies they use in dealing with items. However, no comprehensive study so far seems to have been done regarding how high and low ability students differ in dealing with different formats. Such research is necessary because it might very likely be that a test's structure is unidimensional for one ability group and multidimensional for another group, despite the fact that the tests are unidimensional for the total group (Dorans & Schmitt, 1991).

CHAPTER SUMMARY

In this chapter, studies that compared different item formats in different domains were reviewed. The literature relating to the question of whether MC and CR items can function similarly in tapping students' cognitive ability was explored. In the writing domain, the three studies reviewed consistently showed that there were two distinct characteristics related to MC and CR formats. Three of the four studies reviewed in the verbal domain consistently indicated that format effect did not exist, Janssen and De Boeck (1996) suggested that there was a response-production factor in the free-response synonym task (CR items). The contradictory results from the eight studies reviewed in the quantitative domain make it impossible to conclude whether MC and CR items measure the same thing. Contradictory results were also found from the review in the science domain. It seemed that those different methods and different sub-domains were the main reasons for the different conclusions. Future research is needed to compare item formats across sub-domains in science by applying multiple methods.

Some methodologies have their own limitations, which might affect the results. For example, linear factor analysis, which was the most commonly used methodology in the previous studies, has its weakness with dichotomous data using tetrachoric correlation. Unidimensional

IRT has its limitation in dealing with the local dependence or multidimensionality that might exist in the data.

There are two common limitations across the studies reviewed: (1) only one analytic method was used to address the research question; and (2) few researchers applied cognitive demand analysis to clarify the nature of the differences between formats.

Due to the limitations of the methodologies used in the previous studies, it is important to make an effort to apply other better methodology if possible. In addition, multiple methods (e.g., exploratory and confirmatory approaches) might help to provide stronger evidence to solve this controversial problem. Further, cognitive demand analysis of items may help to clarify the nature of the differences, if any, between items. Finally, it is interesting to find in the literature that high ability students differ from low ability students in both knowledge organization and problem solving strategies. However, so far, no comprehensive research study has been found on how the two groups differ in dealing with different item formats.

CHAPTER III: METHODS AND PROCEDURES

This chapter describes three main research questions, data sources, and test analysis models, as well as their respective procedures. As shown in the previous chapter, studies on format comparison have been done across different domains using different methodologies. However, studies comparing the differences between MC and CR formats in mathematics are lacking, and the methodologies applied (e.g., factor analysis) have shortcomings that might have affected the results. In the present study, two multidimensional item response models (FIFA and MRCMLM) are applied in order to investigate the format differences in the mathematics domain. Four data sets are used to address the three research questions and hypotheses.

RESEARCH QUESTIONS AND HYPOTHESES

1. How do item format factors affect the structure of mathematics tests?

   Hypothesis One: The test structure is two-dimensional, with MC and CR items loading on separate factors.

2. How do MC and CR formats differ in measuring students' cognitive ability beyond knowledge level in mathematics?

   Hypothesis One: The test structure is two-dimensional, with MC and CR items beyond knowledge level loading on separate factors. All MC items load on the first dimension, and all CR items load on the second dimension.

   Hypothesis Two: The test structure is two-dimensional, with one dimension corresponding to overall mathematics proficiency and the other dimension corresponding to CR proficiency. All MC and CR items that measure overall mathematical proficiency beyond knowledge level load on the first dimension, and all CR items that measure CR proficiency beyond knowledge level load on the second dimension.

3. How do differences in item format affect students' performance at different ability levels?

   Hypothesis One: MC and CR item types form two dimensions for low ability students.

   Hypothesis Two: MC and CR item types form one dimension for high ability students.

   Hypothesis Three: There is a statistically significant interaction between item type and ability level.

DATA SOURCES

To address the research questions in the study, data were selected from the data banks of the Third International Mathematics and Science Survey (TIMSS, 1995) and the British Columbia Grade 12 Provincial Examination (1998). TIMSS is the largest and most ambitious study conducted by the International Association for the Evaluation of Educational Achievement for measuring students' mathematics and science achievement in high school curricula across 41 countries. The BC Provincial Mathematics Examination is administered twice each year in British Columbia for Grade 12 students intending to graduate from high school. The two tests, administered in April and August, are designed to be parallel so that those students can take either one of them depending on their schedule.

Two separate analyses across two grade levels (Grade 4 and Grade 8) were done using the TIMSS data bank (Canada). The other two analyses were conducted using the British Columbia Grade 12 Provincial Mathematics Examination data across two time points (April and August 1998).

Third International Mathematics and Science Survey (TIMSS)

A multinational comparative study, TIMSS, was designed to contribute new knowledge about the content of mathematics and science curricula, about how mathematics and science are taught and by whom, and about the outcomes of that teaching reflected in students' achievement and attitudes. The amount of data collected by TIMSS is voluminous in the history of educational research—three population groups, two curriculum areas, 45 countries, a range of grade levels, including curriculum data, achievement measures, and an extensive array of contextual information about educational systems, students, schools, teachers, and instruction.

To address the research questions in the present study, the TIMSS mathematics achievement test was used because it provided a rich source of data for the investigation. There were 26 different clusters of test items assembled into eight booklets. Each student completed one booklet. To select samples that represented each country, the TIMSS developers used two-stage cluster sampling. Schools served as the first stage of selection and classrooms within schools served as the second stage of selection, so that each eligible student had equal probability of being selected. About 17,000 students in each grade level participated in TIMSS across five provinces in Canada. About 2,000 students completed the same test booklet.

Assessment specialists and curriculum experts worked together to ensure that the items used in the tests were appropriate for the students and reflected the curriculum. The TIMSS curriculum framework for mathematics (TIMSS Technical Report, Vol. 1, 1996) indicated that the performance expectations were (1) knowing; (2) using routine procedures; (3) using more complex procedures; (4) mathematical reasoning; and (5) communicating. Different types of achievement items were included in TIMSS. The MC items consisted of a stem and either four or five answer choices. In the CR items, students were asked to construct their own responses to the test questions by writing or drawing their answers. These included both short-answer items and items where students were asked to provide extended responses. The correct response for MC items was scored 1, and 0 for the incorrect response. For some CR items, 0 was assigned for the wrong answer, a partially correct answer was assigned 1, and the correct answer was assigned 2. For other CR items, the wrong answer was assigned 0, and the correct response was assigned 1. The sample and item descriptions of the two analyses using TIMSS data are described as follows.

Data Set One: Grade 3 and Grade 4

In Data Set One, the sample included 2,100 Grade 3 and Grade 4 Canadian students who completed the same booklet (Booklet 5) in TIMSS mathematics examination. High ability students selected were at or above the 70th percentile of the total scores of all the mathematics items in the booklet; low ability students selected were at or below the 30th percentile of the total scores. The reason for selecting the top and bottom 30% of students was to identify clearly different ability groups.

In the mathematics test for Grade 3 and Grade 4 in TIMSS, there are six sub-domains: (1) geometry; (2) measurement, estimation, and number sense; (3) whole number; (4) patterns, relations, and functions; (5) fraction and proportionality; and 6) data representation, analysis, and probability. In Booklet 5 (Grade 3 and Grade 4), there are 40 mathematics items (27 MC and 13 CR items) in all. Only 1 CR item was used in the sub-domains of geometry, data representation, analysis, and probability, and patterns, relations, and functions. Therefore, in order to compare MC and CR items with balanced numbers of items for each format, twenty-nine items were selected for the investigation from the following three sub-domains (1) fractions and proportionality; (2) measurement, estimation, and number sense; and (3) whole number.

TIMSS was based on the framework of Bloom's taxonomy. To avoid the complexity of hierarchical structure of the five levels (knowing, routine procedure, complex procedure, reasoning, and communicating), three distinct hierarchical levels were created by the researcher: (1) knowing and performing routine procedures; (2) performing complex operational procedures; and (3) performing higher mental processes (problem solving, reasoning, or communicating). Because many researchers didn't agree with the hierarchical structure of Bloom's taxonomy involving six levels (Travers, 1980), it was simpler to use the above three clustered levels that were hierarchically distinct for ease of interpretation.

There were 29 items across the three sub-domains in Booklet 5 (see Table 3-1). Three out of 12 CR items were scored 0 for the wrong answer, 1 for partial correct, and 2 for the correct answer. The rest were scored 0 for wrong answer and 1 for the correct answer. The FIFA can only handle responses of 0 or 1, so the three-level responses of CR items were dichotomized as 0 or 1. According to Bock, Gibbons, Schilling, Muraki, Wilson, and Wood (1999), such a procedure works well if the sample size is large, which is the case in the present study.

Table 3-1

Selection of Item Types across Content Area and Level of Cognition of Data Set One

Level of Cognition	Fraction and Proportionality	Measurement, Estimation and Number Sense	Whole Number
Knowledge (K)	2MC + 1CR	2MC + 1CR	4MC + 3CR
Complex Procedure (C)	2MC	2MC	2MC
Higher Mental Process (H)	1MC + 4CR	2MC + 1CR	2CR

Note. 17 MC and 12 CR mathematics items were selected from TIMSS, Grade 3 and Grade 4, Booklet 5. MC = multiple-choice items; CR = constructed-response items.

Data Set Two: Grade 7 and Grade 8

A total of 2,073 Grade 7 and Grade 8 Canadian students completed the same booklet (Booklet 3) in mathematics. As in Data Set One, high ability students were at or above the 70th percentile of the total score, and low ability students were at or below the 30th percentile of the total score in Booklet 3.

In the mathematics test for Grade 7 and Grade 8 (Booklet 3, TIMSS), there were six sub-domains: (1) algebra; (2) measurement; (3) proportionality; (4) fractions and number sense; (5) geometry; and (6) data representation, analysis, and probability. In Booklet 3 (Grade 7 and 8), there were 41 mathematics items (31 MC and 10 CR) in total. Only 2 CR items were used from the sub-domains of geometry, data representation, analysis and probability, and proportionality. To compare MC and CR with balanced number of items in each format, the analysis was focused on the items from the following three sub-domains: (1) fraction and number sense; (2) algebra; and (3) measurement.

The five levels of ability were clustered into three levels: (1) knowing and performing routine procedures; (2) performing complex operational procedures; and (3) performing higher mental processes (problem solving, reasoning, or communicating).

All the items across the three sub-domains and cognitive levels in Booklet 3 (Grade 7 and 8, TIMSS) were used for the analysis in the second data set. There were 26 items in total across the three sub-domains in the booklet; eighteen of them were MC items and 8 of them were CR items (see Table 3-2). Three out of 8 CR items were scored as 0 for the wrong answer, 1 for the partially correct answer, and 2 for the correct answer. The rest of CR items were scored as either 0 or 1. The three-level responses of CR items were dichotomized as 0 or 1.

Table 3-2

Selection of Item Types across Content Area and Level of Cognition of Data Set Two

Level of Cognition	Fraction and Number Sense	Algebra	Measurement
Knowledge (K)	3MC	4MC	3MC
Complex Procedure (C)	3MC	-	1MC
Higher Mental Process (H)	3MC + 3CR	1MC + 2CR	3CR

Note. 18 MC and 8 CR items were selected from TIMSS, Grade 7 and Grade 8, Booklet 5.
MC = multiple-choice items; CR = constructed-response items.

British Columbia Provincial Mathematics Examination

The British Columbia Grade 12 Mathematics Examination is based on Applications of Mathematics 12, which is one of the two courses available across the province. The practical and contextual focus of the course enables students to develop their mathematical knowledge, skill, and attitudes in the context of their lives and possible careers (British Columbia Provincial Examination Specification Document, 1998). The provincial mathematics examination represents 40% of the final letter grade awarded to the student. The provincial examination is divided into two parts: MC questions worth 64% of the total score, and written responses (CR) worth 36% of the total score. The two mathematics examinations (April and August) available each year are equivalent examinations so that students can make the decision as to which examination to take, depending on their schedules. Therefore, the students who take the April examination and those who take the August examination are very similar in terms of their characteristics and academic performances.

According to the test specifications produced by British Columbia Ministry of Education, Skills, and Training (April, 1998), the examination is based on a modified version of Bloom's taxonomy (Bloom et al., 1956). The modified version includes three cognitive levels: knowledge, understanding and application, and higher mental processes. Knowledge questions emphasize the recognition or recall of terminology, specific facts, conventions, classifications, and notations. Understanding and application may require students to form and solve equations, manipulate expressions, or produce a diagram. Higher mental processes include processes of analysis, synthesis, and evaluation. The sample and item descriptions of the two analyses using the British Columbia provincial examination data are described as follows.

Data Set Three: Grade 12, April 1998

A total of 1,718 Grade 12 students in British Columbia completed the Application of Mathematics 12 Examination held in April 1998. High ability students selected were at or above the 60th percentile of the total score, and low ability students selected were at or below the 40th percentile. Due to the different sample sizes in the four data sets (2011, 2073, 1718, and 1429), higher percentage of students (40% at top and bottom) were selected for data sets 3 and 4 so that total numbers of students in each group were approximately the same across the four data sets.

Twenty-three items in total were selected for the analysis. These were from three sub-domains: trigonometry and quadratic relations; exponential logarithmic and polynomial functions; sequences, series and calculus. Eighteen items were MC items and 5 were CR items (see Table 3-3). The 5 CR items were scored at different levels (5, 6, or 7). As in the other analyses, the responses of CR items were dichotomized as 0 or 1.

Table 3-3

Selection of Item Types across Content Area and Level of Cognition of Data Set Three

Level of Cognition	Trigonometry Quadratic Relations	Exponential Logarithmic, Polynomial Functions	Sequences Series Calculus
Knowledge (K)	2MC	2MC	2MC
Complex Procedure (C)	2MC + 2CR	2MC + 2CR	2MC + 1CR
Higher Mental Process (H)	2MC	2MC	2MC

Note. 18 MC and 5CR items were selected from BC Grade 12, April 1998
MC = multiple-choice items; CR = constructed-response items.

Data Set Four: Grade 12, August 1998

A total of 1,429 Grade12 students in British Columbia completed the Application of Mathematics 12 Examination held in August 1998. In total, there were 58 items (50 MC and 8 CR items). High ability students selected were at or above the 60th percentile of the total score of the test, whereas low ability students selected were at or below the 40th percentile.

There were 23 items in total for the three sub-domains: trigonometry and quadratic relations; exponential logarithmic and polynomial functions; sequences, series and calculus. Seventeen of them were MC items and 6 of them were CR items (see Table 3-4). The 6 CR items were scored at different levels (5, 6, or 7). The responses of CR items were dichotomized as 0 or 1.

Table 3-4

Selection of Item Types across Content Area and Level of Cognition of Data Set Four

Level of Cognition	Trigonometry Quadratic Relations	Exponential Logarithmic, Polynomial Functions	Sequences Series Calculus
Knowledge (K)	2MC	2MC	2MC
Complex Procedure (C)	2MC+2CR	2MC + 1CR	2MC + 3CR
Higher Mental Process (H)	1MC	2MC	2MC

Note. 17 MC and 6 CR items were selected from BC Provincial Mathematics Examination, August 1998. MC = multiple-choice items; CR = constructed-response items.

Comparison of Data Sets

The four analyses are different in terms of the following four aspects. First, the data in the study are different because the examinees belong to different grade groups: Grades 3 and 4, Grades 7 and 8, and Grade 12. Second, the four analyses were done on samples from two large data sets: The Third International Mathematics and Science Survey (TIMSS, 1995), and the British Columbia Grade 12 Provincial Mathematics Examination (1998). Although they were both curriculum-based tests, the purposes are different. TIMSS was designed to assess the achievement of high school students for international comparisons, whereas the British Columbia Provincial Examination was designed for high school graduation certification. Third, the test designs of TIMSS and the British Columbia Provincial Examination are different. The tests of TIMSS were expert-made tests, which were pre-tested many times. The participating students were selected based on a two-stage cluster sampling design so that the students selected were representative of the country and the province. However, the British Columbia Department of Education is responsible for the Grade 12 Examinations. The data were the responses from those students who took the examination and then graduated in that year. Because provincial examinations were tests designed for certification purpose, they were not pre-tested. Fourth, tests were conducted at different time points. TIMSS was conducted in April 1995, and the British Columbia Provincial Examinations were held in April and August of 1998.

ANALYSIS MODELS

Full-Information Item Factor Analysis

To test the hypotheses, FIFA was applied as an exploratory approach to see if format was the determining factor for the multidimensionality. In recent years, FIFA (Bock et al., 1988) has often been used for detecting dimensionality (see Zwick, 1987; Carlson & Jirele, 1992; Carlson, 1993). According to Bock et al. (1999), a two- or three-parameter IRT-based FIFA permits a more comprehensive and detailed examination of item dimensionality than is currently available with any other procedure.

The popular method of linear factor analysis based on Pearson (phi) correlation is found not to be able to yield a correct representation of the dimensionality on item pool of dichotomous items (Carroll, 1983; Hulin, Drasgow & Pearsons, 1983; Mislevy, 1986). It is also not satisfactory if the tetrachoric correlation is used for the factor analysis, because such factor analysis can produce spurious factors when items can be answered correctly through guessing (Carroll, 1983; Hulin et al., 1983). Three nonlinear factor analysis models have been proposed to address measurement related issues to overcome the weakness of linear factor analysis. They are McDonald's (1967; 1982) polynomial approximation to a normal ogive model, Christoffersson's (1975) and Muthen's (1978) factor analytic model, and Bock and Altkin's (1981) full-information factor analytic model. In addition to the parametric model, Stout (1987) proposed a non-parametric model based on the covariances of item-pair responses conditioned on extreme trait levels to assess dimensionality.

The FIFA model based on item response theory (Bock & Aitkin, 1981) is regarded as the most sensitive and informative among various methods of investigating the dimensionality of item sets (Bock et al., 1988). FIFA operates on the set of distinct item response vectors instead of on computing correlation coefficients. This method maximizes the likelihood of the item factor loadings and standardized difficulties given the observed patterns of correct and incorrect responses. It solves the corresponding likelihood equations by integrating over the latent distribution of factor scores assumed for the population of examinees. The estimation method is called marginal maximum-likelihood (Bock & Aitkin, 1981), which estimates item parameters using the 1- and 2-parameter normal ogive item-response models. The iterative solution is based on the EM algorithm of Dempster, Laird, and Rubin (1977). In addition, the marginal maximum-likelihood method allows a rigorous test of the statistical significance of factors added successively to the model, which is lacking in the other approaches used for dimensionality testing. The problem of a "not positive-definite" matrix that occurs in tetrachoric correlations can be avoided by listing and saving a "smoothed" positive-definite item-correlation matrix in the item factor analysis procedure (Bock et al., 1999).

Bock and Aitkin proposed the following, based on the m-factor model (for dichotomous data):

That an unobserved response $\underline{y}_{ij}$ for person $\underline{i}$ to item $\underline{j}$ is a linear function of $\underline{m}$ normally distributed latent variables $\theta_i = (\theta_{1i}, \theta_{2i}, \ldots, \theta_{mi})$ and factor loadings $\alpha_j = (\alpha_{j1}, \alpha_{j2}, \ldots, \alpha_{jm})$. This latent response, $\underline{y}_{ij}$ is related to the binary (observed) item response $\underline{x}_{ij}$ through a threshold parameter, $\underline{\gamma}_j$ for item $\underline{j}$, in the following fashion:

$$\text{if } Y_{ij} >= \gamma_j, \text{ then } X_{ij} = 1,$$
$$\text{if } Y_{ij} < \gamma_j, \text{ then } X_{ij} = 0.$$

The probability that examinee $\underline{i}$ with abilities $\theta_i = (\theta_{1i}, \theta_{2i}, \ldots, \theta_{mi})$ will correctly answer item $\underline{j}$ is given by the function:

$$P(x_{ij} = 1/\theta) = \frac{1}{(2\pi)^{1/2}\sigma}\int_{\gamma_j}^{\infty} \exp\left[-\frac{1}{2}\left[\frac{y_{ij} - \sum_{k}^{m} \alpha_{jk}\theta_{ki}}{\sigma_j}\right]^2\right] dy_{ij} \qquad (1)$$

where

$$\sigma = \sqrt{1 - \sum_{k=1}^{m} \sigma_{jk}^{\ 2}}$$

$$\alpha_{jk} = \text{factor loadings}$$

Although the guessing parameter can be included in the FIFA, it is not included as a parameter for MC items in the present investigation. Based on the literature review by Hattie (1984), the three-parameter model makes no difference in detecting dimensionality compared to the two-parameter model. Therefore, the FIFA model (two-parameter) without guessing was applied.

Criteria for FIFA

TESTFACT Software provides a powerful data analytic tool in the form of item factor analysis. The percentage of total variance, root (eigenvalue) criteria, pattern of loadings, and test of statistical significance of factors were examined to determine the dimensionality of the test.

The percentage of total variance has been used to determine whether a test is unidimensional. It is believed that the larger the amount of variance explained by the first component, the closer the set of items is to being unidimensional. Carmines and Zeller (1979) claimed that at least 40% of the total variance should be explained by the first factor in order to be sure about the unidimensionality. Reckase (1979) recommended that the first component should account for at least 20% of the variance. Hattie (1985) pointed out that it is possible for a multidimensional set of items to have a higher variance on the first component than does a unidimensional set. In the present investigation, the criterion by Reckase (1979) was followed as one of the criteria to determine unidimensionality.

An eigenvalue greater than one has been a widely used criterion to determine the number of components to be retained (Kaiser, 1970) in principal component analysis. Many critics (Gorsuch, 1974; Hattie, 1984) have challenged the criterion. Hattie concluded that the criterion resulted in an overestimation of the number of factors in cases where it was known that there was only one factor, and underestimation in the other instances. Therefore, it seems that using the number of eigenvalues greater than one to estimate the number of factors appears to lack justification. In another study, the value of 1.4 was chosen to determine the presence of a second factor based on their simulation results (Smith & Miao, 1994). In the present study, eigenvalue (root value) of 1.4 was used to determine the presence of the significant factor.

The Promax method (Hendrickson & White, 1964) was chosen because it is especially appropriate for item analysis. The method allows for identifying clusters of items that form unidimensional subsets within a heterogeneous collection of items (Bock, et al., 1999). It can be done by first rotating to an orthogonal varimax solution and then relaxing the orthogonality of the factors to better fit simple structure. From the orthogonally rotated matrix, an ideal factor matrix is constructed in which high loadings of varimax are made higher and low loadings smaller. Such a procedure is accomplished by normalizing the orthogonal matrix by rows and columns and taking the k^{th} power of each loading. This final step is to find the least squares fit to the ideal matrix, using Hurley and Cattell's (1962) Procrustes technique. This best fit is then the oblique solution. In terms of the value of the k^{th} power to which the varimax loadings are raised, Hendrickson and White concluded that the optimal value of k is 4; however, for the occasional factor analysis where the data is particularly cleanly structured, a lower power seems to provide the best solution. In the TESTFACT program, a default value of 4 is used.

The chi-square difference test was applied to see if an additional dimension improved the model fit. If the sample size is large, a test of the fit of the multiple factor models can be obtained using a chi-square approximation to the likelihood ratio test.

The likelihood-ratio chi-square statistic is defined as,

$$G^2 = 2\sum r_l In\left(\frac{r_l}{N\tilde{P}_l}\right) \qquad (2)$$

where

$\underline{r}_l$ is the frequency of response vector $\underline{l}$,

$\underline{P}_l$ is the probability of response vector $\underline{l}$,

The degrees of freedom for these statistics are equal to,

$$2^n\,(m + 1) + m\,(m - 1)\,/2,$$

where $\underline{n}$ is the number of items,

$\underline{m}$ is the number of factors.

In order to access whether the additional dimension is better, a chi-square difference test was conducted. The G^2 difference test was computed in the following fashion,

$$G^2{}_{diff} = G^2{}_{1\text{-}F} - G^2{}_{2\text{-}F} \qquad (3)$$

where

$\underline{G^2{}_{1\text{-}F}}$ is the value of the likelihood-ratio chi-square statistics obtained after fitting a one-factor model, and

$\underline{G^2{}_{2\text{-}F}}$ is the value of the likelihood ratio chi-square statistic obtained after fitting a two-factor model.

The degrees of freedom for the difference test were also computed by subtracting those associated with the one- and two-factor model fit statistics. The model can be re-estimated and the test can be repeated for successive values of the factors. The resulting difference between these statistics is also distributed as chi-square. However, it has been pointed out that the statistics must be interpreted with caution (Bock et al., 1999). The results of previous research have indicated that the reliance on the fit statistics alone may lead to overfactoring (Zwick, 1987). The fit statistics were unable to correctly identify the number of dimensions based on a simulation study with a small number of replications (Berger & Knol, 1990). Champlain and Gessaroli (1998) assessed the dimensionality using the fit statistics based on FIFA. For a two-factor simulation study, the number of false acceptances of unidimensionality was significantly greater for data sets in which a correlation of .70 was specified between latent traits (138 out of

1500). However, there was no false acceptance of unidimensionality for those data sets in which a zero correlation was specified between latent traits.

According to Zwick (1987), the size of eigenvalues, percentage of variance explained, and the pattern of loadings should also be considered with the fit statistics in determining the number of factors. In the present study, the above three criteria were considered together in determining the number of factors.

Multidimensional Random Coefficient Multinomial Logit Model

The Multidimensional Rasch model (Multidimensional Random Coefficients Multinomial Logit Model, MRCMLM) was used to test if the two different item types (MC vs. CR) measured two different latent proficiencies (constructs). The comparison between multidimensional and unidimensional item response models might help to reveal whether MC and CR item types represent different latent proficiencies. One limitation of the commonly used item response models is their unidimensionality assumption, which means that the test measures only a single latent trait (Hambleton & Murray, 1983; Lord, 1980). In reality, many tests are composed of subdomains that indicate that there are multiple traits being assessed (Reckase, 1979; Traub, 1983). The use of unidimensional item response theory will bias parameter and ability estimation if the underlying dimensions are not highly correlated (Folk & Green, 1989).

The rationale for applying multidimensional models in a format comparison study is that the complexity of tests in the real world might not meet the assumption of unidimensionality of the item response models (Hambleton & Swaminathan, 1985). In previous studies, many researchers based their conclusions on a single model (e.g., linear factor analysis, unidimensional IRT model). Because of the complexity of the problems pervasive in different models, it might be useful to apply multidimensional models. In addition, both exploratory (FIFA) and confirmatory (MRCMLM) approaches are highly desirable (Joreskog, 1974) because they might provide richer evidence of the differences between MC and CR items.

The MRCMLM was developed to address the problems of unidimensional and multidimensional IRT models. One limitation of IRT is its unidimensionality assumption, which means that the test measures only a single latent trait (Hambleton & Murray, 1983; Lord, 1980). Recently, a multidimensional IRT model has been developed and studied in order to overcome the drawbacks of unidimensional IRT models (Ackerman, 1992; Anderson, 1985; Camili, 1992).

MRCMLM was developed to provide researchers a great deal of flexibility to form customized models as well as to allow the analysis of polytomously scored data. In the present study, MRCMLM was applied as a confirmatory approach to address the research questions of whether MC and CR items are similar in testing students' cognitive ability in mathematics. Items were forced to be indicators of certain latent traits based on the research hypotheses.

The MRCMLM assumes that a set of D traits underlie a person's responses. The probability of a response in category k of item i is modeled as,

$$\Pr\left(\mathbf{X}_{ik} = 1; \mathbf{A}, \mathbf{B}, \xi / \theta\right) = \frac{\exp\left(\mathbf{b}'_{ik}\theta + \mathbf{a}'_{ik}\xi\right)}{\sum_{k=1}^{k_i} \exp\left(\mathbf{b}'_{ik}\theta + \mathbf{a}'_{ik}\xi\right)} \qquad (4)$$

where

Pr ($\mathbf{X}_{ik} = 1$; $\mathbf{A}$, $\mathbf{B}$, ξ/θ) is the probability of a response in category k of item i,

$\mathbf{b'}_{ik}$ is a vector of the scoring function of response k to item i,

ξ is a vector of p free item parameters (e.g., difficulty, bias),

$\mathbf{a'}_{ik}$ is a design vector, a linear combination of ξ for response category k in item i,

$\mathbf{x}_{ik}$ is the response to item i.

There are three submodels (Wang, 1994) that are very useful in practice: multidimensional between-item models, multidimensional within-item models, and multidimensional mixed models. In the present study, only the multidimensional between-item model and within-item model were used to address the research questions.

Multidimensional Between-Item Model

If tests are made up of sub-sets of items that are mutually exclusive and measure different latent variables, then the multidimensional between-item model can be applied. Each item on the test serves as an indicator for a single latent dimension as shown in Figure 1. In the present study, the between-item model was applied in a confirmatory approach to test the hypothesis that the test might be multidimensional due to item format, sub-content, or cognitive factors.

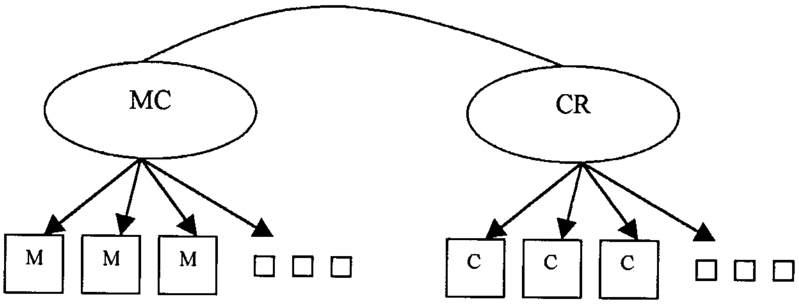


Figure 1. Two-Dimension Between-Item Format Model

M—MC items, C —CR items, □ —other items

Multidimensional Within-Item Model

If each of the items in a test can be an indicator of multiple latent dimensions, then a within-item model approach can be applied. In the present investigation, it was hypothesized that CR item types had some unique characteristics that differed from the overall mathematics ability that test developers intended to measure. Therefore, each CR item was an indicator of both latent dimensions (overall mathematics ability vs. CR proficiency, as shown in Figure 2). Finally, the within-item model was compared with the unidimensional model as well as with the between-item model to see which model fit the data best.

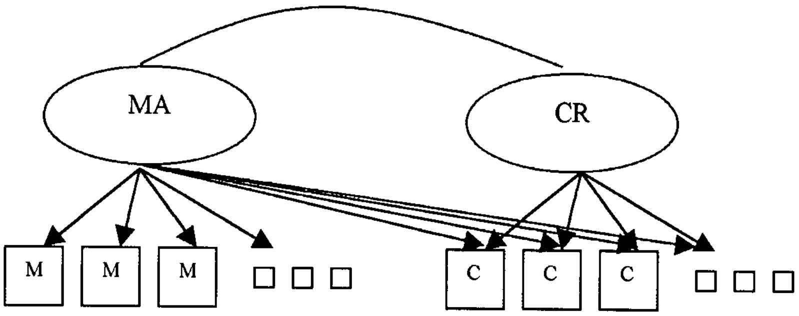


Figure 2. Two-Dimension Within-Item Format Model.

MA—Mathematics Proficiency, M—MC items, C—CR items, ...—other items.

Wang (1994) compared the between-item model and the two other approaches. His study indicated that the IRT approach was practically preferred only when the sub-scales were very highly correlated or when the dimensions were theoretically indistinguishable.

According to Adams et al. (1997), under the confirmatory factor analysis (consecutive) approach, if the underlying dimensions were not orthogonal, the estimation accuracy of the consecutive approach was lower than that of the multidimensional IRT approach for the item parameters when the dimensions were estimated separately rather than simultaneously. Chang and Davison (1992) found that both the bias and standard errors for subtests were smaller when a multidimensional approach was applied.

Criteria for MRCMLM

The computer program ConQuest developed by Wu, Adams, Wilson (1997) was designed to test a variety of IRT models including multidimensional models. ConQuest produces marginal maximum likelihood estimates for the parameters of the models. The estimation algorithms used are adaptations of the quadrature method described by Bock and Aitken (1981) and the Monte Carlo method of Volodin and Adams (1995). The number of nodes for each dimension used in the study is 20, which is recommended as safe by Wilson and Adams (1993). The higher number of nodes would result more accurate estimation of parameters. The iteration criteria used is 0.001, which means that the EM algorithm is terminated when the largest absolute change in any parameter estimate becomes less than 0.001 (default). Further investigation seems to be needed in terms of the rule of convergence according to Adams, Wilson and Wang (1997).

The fit of the models is ascertained by generalizations of the Wright and Masters' (1982) residual-based methods that were developed by Wu (1997). If $\underline{A}_{\underline{p}}$ is the $\underline{p}^{\text{th}}$ column of the design matrix $\underline{A}$, then the Wu fit statistic is based upon the standardized residual,

$$z_{np}(\theta_n) = \left(A'_p x_n - E_{np}\right) / \sqrt{V_{np}} \qquad (5)$$

where

$\underline{A'_p x_n}$ is the contribution of person $\underline{n}$ to the sufficient statistic for parameter $\underline{p}$,

$\underline{E_{np}}$ is conditional expectation of $\underline{A'_p x_n}$,

$\underline{V_{np}}$ is conditional variance of $\underline{A'_p x_n}$.

A weighted fit statistic (each squared residual is weighed by its variance) was used in the study because responses made by persons for whom the item is remote have less influence on the magnitude of the item fit statistics (Wright and Masters, 1982, p.101). A disadvantage of the unweighted procedure is that it is rather sensitive to unexpected responses made by persons for whom item $\underline{i}$ is far too easy or far too difficult.

To construct a weighted fit statistics, the square of this residual is averaged over the cases and then integrated over posterior ability distributions. In ConQuest, the Monte Carlo method is used to approximate the integrals. Wu (1997) has shown that such statistics have approximate scaled chi-squared distributions. These statistics are transformed to approximate normal deviates using the Wilson-Hilferty transformations. The derivation and justification for these transformations was given in Wu (1997).

In the present study, an item with a $\underline{z}$ value less than 2.0 was regarded as a fit item. If an item had a $\underline{z}$ value between 2.0 to 4.0, it was regarded as a misfit item. If an item had a $\underline{z}$ value larger than 4.0, it was regarded as a serious misfit item. However, according to Hambleton and Murray's simulation study (1983), the sample size can significantly impact the detection of misfittng items. Based on their simulation study, misfit items increased from 5 to 38 out of 50 items when sample size increased from 150 to 2400. It seems that sample size around 600 to 1000 may give accurate results. Therefore, the results in the present study were analyzed and interpreted cautiously when sample size was over 1000.

The comparison between different models was based on the difference of deviances—a -2 log likelihood statistic that indicates how well the item response model fit the data. The difference between the deviances follows the chi-square distribution asymptotically. Whether one model is significantly (statistically) better than the other model can be determined, but large chi-square values may appear due to the large sample size (Hambleton & Swaminathan, 1985). Therefore, the results in the present study should be analyzed and interpreted cautiously when sample size is over 1000.

In the previous studies, researchers used the one-parameter IRT-based indices to detect dimensionality because they regarded such indices as the most direct tests of unidimensionality in the calibration process. They also believed that there is evidence that a given set of items refers to the unidimensional ability if it fits the Rasch model. However, indices based on the one-parameter IRT model were criticized due to its insensitivity to violations of unidimensionality (Wallenberg, 1982; and Rogers, 1984). In Reckase (1979), Rasch fit statistics were highly correlated with guessing, indicating that guessing was the major component in those statistics instead of discrimination or multidimensionality. Rogers (1984) investigated the performances of all those indices that were based on one-, two-, or three- parameters, and concluded that there was an increased sensitivity to multidimensionality as more parameters were fitted. Hattie (1984, 1985) did a comprehensive literature review as well as a simulation study to investigate various indices that were related to dimensionality testing. The results from these studies suggested that those indices based on the size of residuals after fitting a two- or three-parameter latent trait model were the most useful in detecting unidimensionality.

In the present study, the confirmatory method of comparing the residual differences of one- and two-dimensional IRT models was a better approach for detecting multidimensionality than the unidimensional IRT model. The advantages of using MRCMLM are (1) that it is possible to confirm or reject the hypothesis of multidimensionality statistically; and (2) the correlation between the two latent traits can be found.

Summary

In order to test the research hypotheses, the results from the exploratory factor analysis (FIFA) were examined first, and then the confirmatory factor analysis approach based on MRCMLM was used to see if the conclusions based on the results yielded by FIFA could be supported.

ANALYSIS PROCEDURES

The detailed procedures for testing the hypotheses are described in this section. There were seven stages in the analyses and the same procedure was applied in the analyses of each data set.

Stage One: Recoding and Information Loss

In order to apply FIFA that can only handle dichotomous response type, the multilevel responses of the CR items were dichotomized as 0 or 1. For some CR items, they were coded as 0 or 1, which is similar to MC items. However, for some others, 0 was for a wrong answer, 1 was for a partially correct answer, and 2 was for the correct answer. Sometimes, CR items had more levels for partially correct answers (e.g., 0, 1, and 2). For the two-point open-ended response items, scores of zero, one were re-coded to a value of zero, while score of two was re-coded to a value of one. For the four-point open-ended response items, scores of zero, one, and two were re-coded to a value of zero, while scores of thee and four were re-coded to a value of one.

The information loss due to this re-coding was examined using the IRT test information index and plot. It was hoped that such dichotomization would not lead to too much information loss. Therefore, the item and test information values of the two data sets (original and dichotomized) were compared. The difference of the information value was regarded as loss of information after the dichotomization.

The item and test information values of the two data sets were computed and plotted using FLUX (Burket, 1993). The item information $I_{jk}(\theta)$ is defined by

$$I_{jk}(\theta) = \alpha^2 \sigma^2 (x_j / \theta) P_{jk}(\theta), \qquad (6)$$

$$\sigma^2 (x_j / \theta) = \sum_{k=1}^{m_j} (k-1)^2 P_{jk}(\theta) - \left[E(x_j / \theta)\right]^2, \quad (7)$$

where

$I_{jk}(\theta)$ represents the information for each of the score levels of the jth item and $I_j(\theta)$ is the information for the item as a whole. $I_{j.}$ represents the total test information.

$$I_j(\theta) = \sum_{k=1}^{m_j} I_{jk}(\theta) \qquad (8)$$

$$I_{j.} = \int I_j(\theta) \delta\theta \qquad (9)$$

$$E(x_j / \theta) = \sum_{k=1}^{m_j} (k-1) P_{jk}(\theta). \qquad (10)$$

The test information values of the original data and the dichotomized data were compared to see if the loss was severe. According to Ercikan et al. (1998), less than 5 % of the information loss was regarded as trivial. More than 15 % of the information loss was regarded as severe.

Stage Two: Treatment of Missing and not Attempted

The missing data and not-attempted data in the file were examined to see if there was any speededness in the test. If more than 5 items at the end of a test had more than 10% of the not-attempted responses, then the test can be regarded as having speededness. If there was speededness, the not-attempted responses were ignored, which is the best available procedure

(Lord, 1980). If there was no evidence of speededness, the missing and not-attempted data were coded as 0.

Stage Three: Classical Test Analysis

A detailed description of the data was obtained based on the classical test theory using SPSS. The mean, standard deviation, item difficulty, reliability (Cronbach's alpha), and item discrimination (item total correlation) were obtained. Cronbach's alpha was used because the split-half method tended to be an underestimate of the reliability coefficient of the full-length test (Crocker & Algina, 1986, p.137). The differences between MC and CR items in terms of the above indices were examined as well.

Stage Four: Comparison of Different Models Related to Research Question One

To address the first research question, FIFA and MRCMLM were applied as exploratory and confirmatory approaches, respectively. In the exploratory procedures, three models (one-, two- and three-factor) were compared to see which one was the best model for the data. In the confirmatory procedures, two models (one- and two-dimensions) were compared. The confirmatory procedures were as follows:

(1) A one-dimension model was used to fit the data. In this model, all items were regarded as testing one underlying dimension.

(2) A two-dimension between-item model (Figure 1, see p.40) was used to fit the data. In this model, all MC items were set to load on one dimension and all CR items on a second dimension.

Stage Five: Comparison of Different Models Related to Research Question Two

To address the second research question, items that were designed beyond knowledge level were selected for the investigation. For example, sixteen items designed to tap high cognitive levels (complex procedures and high mental process) in TIMSS Mathematics Examination (Gr.3 and Gr.4) were selected to address the research question two. FIFA and MRCMLM were applied as exploratory and confirmatory approach, respectively. Three models (one-, two- and three-factor) were compared to see which one was the best model for the data. The comparison of the above three models can address the research question in terms of whether MC and CR items beyond knowledge level differed in measuring students' mathematical ability. Similarly, the results from the confirmatory analyses can also reveal whether MC and CR items beyond knowledge level differed in assessing students' mathematical ability. The confirmatory procedures were as follows:

(1) In order to examine whether the MC and CR items beyond knowledge level were measuring the same proficiency, all items were regarded as testing one underlying dimension.

(2) A two-dimension between-item model as shown in Figure 1 (see p. 40) was hypothesized (all MC items measured one dimension and all CR items measured a second dimension).

(3) A two-dimension within-item model as shown in Figure 2 (see p. 40) was hypothesized (all MC and CR items measured overall mathematical proficiency – first dimension and all CR items measured CR proficiency – second dimension). This procedure was to detect whether those CR items measured the proficiency that was different from the overall mathematics proficiency.

Stage Six: Comparison of Different Ability Groups Related to Research Question Three

To address the last research question, similar analyses as in stage five were followed separately for the two groups (high ability vs. low ability). The high ability students selected were at or above the 70th (data sets 1 and 2) and 60th (data sets 3 and 4) percentile of the total score. Low ability students were at or below the 30th (data sets 1 and 2) and 40th percentile (data set 3 and 4). Higher percentage of students (40% at top and bottom) were selected in data sets 3 and 4 because of the smaller sample size (1718, and 1429) compared to that of data sets 1 and 2 (2011, and 2073).

In addition, a repeated measures design (one between factor and one within factor) was used to examine whether the performance of high and low ability students differed when the students responded to MC and CR items. A univariate approach was followed in order to interpret the results. A statistically significant interaction between the format and group variable was hypothesized. The effect size index f recommended by Cohen (1988) was used to judge the strength of the effect.

Cohen suggested that .10, .25, and .40 can be regarded as small, medium, and large effect sizes. The effect size index f is defined as follows:

$$f = \sqrt{\frac{\eta^2}{1-\eta^2}} \qquad (11)$$

where

η is the proportion of the total variance $\sigma_t{}^2$ made up of the variance of the population means $\sigma_m{}^2$: $\eta^2 = (\sigma_m{}^2 / \sigma_t{}^2) = (\sigma_m{}^2 / (\sigma^2 + \sigma_m{}^2))$.

Stage 7: Examination of Factor Loadings

Two researchers examined factor loadings in the present investigation in order to get a clearer picture of the unexpected loading patterns from the exploratory factor analysis. Both researchers worked in a large testing company. One of them had 20 years of experiences as a senior research scientist. The items that failed to load on any factor and the items that loaded on the minor dimensions were examined by looking at the cognitive demand that might be involved. The cognitive demand in mathematics problem solving may involve the following processes (Mayer, 1980), (1) problem representation; (2) schema knowledge; 3) computation skills; and 4) strategic knowledge. In the present investigation, items were examined based on the above model. Both investigators conducted the analyses of items using the same procedures. When the two researchers ended up with different interpretations, further discussions were conducted until the agreement was reached.

CHAPTER SUMMARY

In this chapter, the research questions and hypotheses, data for the four analyses and latent trait models used were described. Three research questions were presented (1) How does the format factor affect the dimensionality of the structure of mathematics test? (2) How do MC and CR formats differ in measuring students' higher cognitive ability in mathematics? (3) How do differences in item format affect students' performance at different ability levels?

Four different data sets were described. The four data sets differed in terms of the grade range, test purpose, test structure, and time points. Such variation can help to generalize the results. Detailed procedures of the analyses at each stage were described as well.

To answer the research questions, two multidimensional item response models (FIFA and MRCMLM) were identified for testing the dimensionality of the test. FIFA was identified as an exploratory approach and MRCMLM as a confirmatory approach. FIFA, a combination of item response theory and factor analysis, was regarded as the most sensitive and informative among various methods of investigating the dimensionality of item sets. It overcomes the weakness of linear factor analysis by operating on the set of distinct item response vectors instead of computing correlation coefficients. MRCMLM, a multidimensional IRT model, can provide researchers with a great deal of flexibility to form customized models for hypothesis testing. It was hypothesized that MC and CR items measured same mathematical ability. If both exploratory and confirmatory factor analysis results agreed, then the hypotheses were strongly supported. The analyses of factor loadings were used to help with the interpretation of the results of the factor analyses. The analyses of cognitive demand of items were based on Mayer's (1980) cognitive process models relating to mathematics problem solving.

CHAPTER IV: RESULTS

EXAMINATION OF DATA SET ONE: TIMSS (GRADE 3 AND GRADE 4)

In this section, the results from the first data set were presented. First of all, detailed results from the classical test theory such as reliability, item difficulty, and item discrimination were presented. Second, results from the two multidimensional IRT models: Full-Information Factor Analysis (FIFA) and Multidimensional Random Coefficient Multinomial Logit Model (MRCMLM) were displayed and discussed. The criteria for judging the dimensionality of the test structure using FIFA was based on the size of roots, the total variance accounted for by the factors, factor loadings and the goodness-of-fit test. The criteria for determining the best models based on MRCMLM were the chi-square test of the deviance difference between models, the mean square residual index, and the standardized fit index (z). Third, the examinations of the cognitive demand of the items were used to help with the interpretation of the results of the factor analyses. Fourth, the differences between high and low ability students were compared in terms of their responses to the different formats using two latent trait models as well as analysis of variance procedures.

Missing Data

The first analyses were based on the data set of TIMSS (Booklet 5, Gr.3 and Gr.4), from which 29 items were selected for the investigation. The Canadian sample (N = 2,011) was selected from the TIMSS data bank. By looking at Table 4-1, it can be seen that students omitted more CR items than MC items: the average response rates for CR and MC items were 92.6% and 97.7%, respectively. The second CR item #2 had the highest missing response rate (7.9%) and the sixth CR item #6 had the highest not-attempted response rate (10.1%). Speededness did not seem to be a problem because no item at the end of the test had more than 10% of missing responses. Therefore, all the missing data and not-attempted responses were coded as 0 in the data file.

Constructed-Response Items

Three CR items originally coded as 0, 1, and 2 were dichotomized to be either 0 or 1. For the two-point open-ended response items, scores of zero and one were re-coded to a value of zero, while score of two was re-coded to a value of one. It was expected that some information was lost. The information function value was computed based on the item estimates using

FLUX, and also the information function was plotted to see if there was any information loss after the 3 CR items were dichotomized. The loss of information was 4%, which was small (see Figure 3, Appendix A). Therefore, it is possible to conclude that dichotomizing the 3 CR items did not lead to loss of information.

Table 4-1

Percentage of Missing Data (Data Set One)

Item	Valid Response Rate	Percent of Missing Response	Percent of not attempted
1. MC1	98.4	1.6	0.0
2. MC2	99.0	1.0	0.0
3. MC3	98.1	1.9	0.0
4. MC4	98.8	1.2	0.0
5. MC5	98.7	1.3	0.0
6. MC6	97.0	3.0	0.0
7. MC7	98.0	2.0	0.0
8. MC8	94.8	5.2	0.0
9. MC9	98.4	1.6	0.0
10. MC10	98.9	1.1	0.0
11. MC11	98.2	1.8	0.0
12. MC12	97.9	2.1	0.0
13. MC13	96.8	3.2	0.0
14. MC14	98.4	1.6	0.0
15. MC15	97.1	2.9	0.0
16. MC16	94.5	5.5	0.0
17. MC17	97.7	2.3	0.0
Mean	97.7	2.3	0.0
18. CR1	100.0	0.0	0.0
19. CR2	89.3	**7.9**	2.8
20. CR3	91.4	4.3	4.3
21. CR4	90.2	4.4	5.5
22. CR5	90.0	3.8	6.2
23. CR6	89.1	0.8	**10.1**
24. CR7	100.0	0.0	0.0
25. CR8	91.5	5.9	2.6
26. CR9	92.7	3.6	3.7
27. CR10	100.0	0.0	0.0
28. CR11	89.3	5.8	5.0
29. CR12	87.4	4.3	8.3
Mean	92.6	3.4	4.04

Note. 29 items were selected from TIMSS, Gr.3 and Gr.4 Mathematics Examination (Booklet5). Sample size is 2,011.

Results from Classical Test Theory

From Table 4-2, it can be seen that students scored higher on MC items than on CR items (item mean score for MC items was 0.54; item mean score for CR items was 0.48). The item total correlation varied from 0.17 to 0.56. On average, CR items had higher item total correlation than MC items. The overall reliability (Cronbach's alpha) was 0.83.

Table 4-2

Descriptive Item Analysis[a] (Data Set One)

Item	Cognitive Level[b]	Sub-content[c]	Item Mean	Std Dev	Item Total Correlation[d]
1. MC1	K	FN	0.75	0.43	0.37
2. MC2	C	FN	0.50	0.50	0.28
3. MC3	K	WN	0.51	0.50	0.41
4. MC4	K	WN	0.83	0.37	0.35
5. MC5	K	ME	0.75	0.43	0.24
6. MC6	C	ME	0.42	0.49	0.31
7. MC7	K	WN	0.77	0.42	0.34
8. MC8	K	FN	0.46	0.50	0.35
9. MC9	H	WN	0.56	0.50	0.17
10. MC10	K	WN	0.55	0.50	0.40
11. MC11	K	ME	0.45	0.50	0.27
12. MC12	C	FN	0.32	0.47	0.24
13. MC13	C	ME	0.50	0.50	0.22
14. MC14	C	WN	0.83	0.38	0.34
15. MC15	H	ME	0.44	0.50	0.21
16. MC16	C	WN	0.33	0.47	0.41
17. MC17	H	ME	0.15	0.36	0.24
18. CR1	H	ME	0.50	0.50	0.35
19. CR2	K	FN	0.54	0.50	0.26
20. CR3	H	FN	0.50	0.50	0.56
21. CR4	H	FN	0.33	0.47	0.54
22. CR5	H	FN	0.73	0.45	0.37
23. CR6	K	WN	0.69	0.46	0.41
24. CR7	H	FN	0.46	0.50	0.29
25. CR8	K	WN	0.38	0.49	0.55
26. CR9	K	WN	0.54	0.50	0.45
27. CR10	H	WN	0.32	0.47	0.27
28. CR11	H	WN	0.42	0.49	0.47
29. CR12	K	ME	0.32	0.47	0.44
Mean for MC			0.54		
Reliability[e]			0.68		
Mean for CR			0.48		
Reliability			0.75		
Mean			0.51		
Reliability for MC and CR			0.83		

Note. [a] 29 items from Booklet 5, Gr.3 and Gr.4, TIMSS (N=2,011). [b].K-knowledge, C-complex procedure, H-higher mental process. [c] FN-fraction and number sense, ME-measurement, WN-whole number. [d] Discrimination index: point-biserial. [e] Reliability-Cronbach's alpha.

RESEARCH QUESTION ONE

Exploratory Factor Analysis

Three models were tested to investigate whether item format factors affect the structure of the mathematics test.

Hypothesis one: The test structure is unidimensional.

Hypothesis two: The test structure is two-dimensional.

Hypothesis three: The test structure is three-dimensional.

FIFA was applied as an exploratory approach to test the three hypotheses. According to the variance and root indices in Table 4-3, it seemed that the one-factor model was the best of all the three models. The total variance accounted for by the model was 26.1%. The largest root was 7.59, which was greater than 1.4, indicating a significant factor. Hypothesis one was supported. In terms of the two-factor model, the first factor was a significant factor ($\lambda = 7.61$), but the second factor was not ($\lambda = 0.75$). Hypothesis two was rejected. Similarly, the first factor was a significant factor ($\lambda = 7.64$) in the three-factor solution, but the second ($\lambda = 0.83$) and third factors ($\lambda = 0.76$) were not significant factors. Hypothesis three was rejected as well.

Examination of Factor Loadings

All the loadings of the one-factor solution were above 0.30 except 3 MC items (#9, #13 and #15) and 1 CR item #2. These four items had medium item difficulty and low item discrimination (see Table 4-2).

MC item#9 Janis ate ½ cake, maija ate ¼, mother ate ¼, how much remains?

A. 1/2 B. 1/4 C. 1/8 D. 0

MC item#13 Which of the following could be the weight (mass) of an adult?

A. 20 kg B. 5 kg C. 75kg D. 900kg

MC item#15 The weight (mass) of a clothespin is 9.2g, which of these is the best estimate of the total weight (mass) of 1000 clothespins?

A. 900g B. 9 000g C. 90 000g D. 900 000g

CR item#2 Write a fraction that is larger than 2/7.

Table 4-3

Comparison of Factor Loadings of the Three Solutions Based on FIFA[a] (Data One)

			One-factor[d]	Two-factor[e]		Three -factor		
Item	CL[b]	SC[c]	Factor 1	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3
1. MC1	K	FN	**0.573**	**0.547**	0.071	**0.535**	-0.024	0.100
2. MC2	C	FN	**0.392**	**0.339**	0.084	0.288	-0.065	0.200
3. MC3	K	WN	**0.564**	**0.533**	0.072	**0.542**	0.064	-0.004
4. MC4	K	WN	**0.593**	**0.471**	0.175	**0.451**	0.100	0.095
5. MC5	K	ME	**0.351**	0.219	0.167	0.230	0.165	-0.009
6. MC6	C	ME	**0.415**	**0.476**	-0.040	**0.536**	0.022	-0.129
7. MC7	K	WN	**0.525**	**0.429**	0.141	**0.426**	0.079	0.064
8. MC8	K	FN	**0.469**	**0.375**	0.135	**0.429**	0.204	-0.125
9. MC9	H	WN	0.240	0.163	0.101	0.158	0.077	0.028
10. MC10	K	WN	**0.553**	**0.681**	-0.100	**0.730**	-0.111	-0.052
11. MC11	K	ME	**0.362**	0.078	**0.339**	0.057	**0.341**	0.027
12. MC12	C	FN	**0.329**	0.230	0.130	0.207	0.065	0.090
13. MC13	C	ME	0.297	0.176	0.154	0.183	0.149	-0.004
14. MC14	C	WN	**0.552**	**0.390**	0.214	**0.345**	0.069	0.191
15. MC15	H	ME	0.283	0.278	0.025	**0.366**	0.169	-0.236
16. MC16	C	WN	**0.580**	**0.720**	-0.117	**0.740**	-0.174	0.023
17. MC17	H	ME	**0.401**	**0.623**	-0.226	**0.637**	**-0.362**	0.107
18. CR1	H	ME	**0.468**	0.246	0.274	0.199	0.181	0.147
19. CR2	K	FN	0.041	-0.163	0.230	-0.133	**0.326**	-0.121
20. CR3	H	FN	**0.774**	**0.422**	**0.436**	0.168	-0.081	**0.802**
21. CR4	H	FN	**0.780**	**0.447**	**0.417**	0.201	-0.088	**0.781**
22. CR5	H	FN	**0.548**	0.000	**0.636**	-0.147	**0.303**	**0.494**
23. CR6	K	WN	**0.594**	0.181	**0.493**	0.099	0.285	0.297
24. CR7	H	FN	**0.386**	0.153	0.283	0.123	0.254	0.066
25. CR8	K	WN	**0.762**	**0.749**	0.063	**0.754**	-0.026	0.074
26. CR9	K	WN	**0.614**	**0.432**	0.240	**0.443**	0.218	0.011
27. CR10	H	WN	**0.371**	-0.113	**0.571**	-0.160	**0.747**	-0.099
28. CR11	H	WN	**0.615**	0.096	**0.626**	0.033	**0.682**	0.036
29. CR12	K	ME	**0.598**	**0.414**	0.241	**0.393**	0.119	0.140
Variance			26.14%	26.69%	2.58%	26.50%	2.58%	1.99%
Largest roots			7.59	7.61	0.75	7.64	0.83	0.76
Correlation								
Factor 1				1		1		
Factor 2			-	0.672	1	0.719	1	
Factor 3						0.712	0.677	1

Note. [a] 29 items from Booklet 5, Gr.3 and Gr.4, TIMSS (N=2,011). [b] CL-cognitive level, K- knowledge, C-complex procedure, H-higher mental process. [c] SC-sub-content, FN-fraction and number sense, ME-measurement, WN-whole number. [d] Factor loadings for one-factor model are principal factor loadings. [e] Factor loadings for two- and three-factor models are promax factor loadings. [f] Bold numbers are loadings greater than the absolute value of 0.30.

The examination of the four items revealed that the following cognitive demand might be put on the students. First, examinees needed to translate the problem into an internal representation (e.g., equation). Second, schema knowledge seemed to be needed. For example, for MC #15, it was important for the examinees to know the equation (total weight = number of clothespins x weight of each clothespin). Third, computation skills (e.g., arithmetic rule) were required to produce the correct answer. For MC #13 and CR #2, the examinees did not have to go through the third step. Although MC #9 and MC #15 required the examinees to go through all the three steps, the computation skills required were very basic. By looking at the other items that had higher loadings on the factor (e.g., CR #3), it seemed that the examinees needed to apply higher level of computation skills in order to get the correct answer.

Two pairs of CR items (#3 and #4, #10 and #11) were found to represent the two non-significant minor dimensions in the three-factor solutions.

CR #3 and CR #4

CR #3. Maria and her sister Louisa leave home at the same time and ride their bicycles to school *9* kilometers away. Maria rides at a rate of *3* kilometers in *10* minutes. How long will it take her to get to school?

CR #4. Louisa rides at a rate of *1* kilometer in *3* minutes. How long will it take her to get to school?

CR #10 and CR #11

In a game, Mysong and Naoki are making addition problems. They each have four cards like these.

$$\boxed{1}\qquad\boxed{2}\qquad\boxed{3}\qquad\boxed{4}$$

The winner of the game is the person who can make the problem with the largest answer.

Mysong placed the cards like this

$$\begin{array}{r} \boxed{4}\;\boxed{3} \\ +\,\boxed{2}\;\boxed{1} \\ \hline \end{array}$$

Naoki placed the cards like this

$$\begin{array}{r} \boxed{3}\;\boxed{1} \\ +\,\boxed{2}\;\boxed{4} \\ \hline \end{array}$$

CR #10 Who won the game? ______________

CR #11 How do you know? ________________________________

CR #3 and CR #4 were complex problem solving tasks. First, students were required to use schema knowledge to represent the structure of the problem. Second, conceptual knowledge (rule) of the mathematical equation of the variables (distance, speed, and time) that matched the schema was needed. Third, computation skill was required to produce the correct answer. The two items were regarded as locally dependent items because the success of one item meant the success of the other item. Similarly, CR #10 and CR #11 were two locally dependent items too. However, the two minor dimensions represented by the two pairs of items seemed to be trivial according to the variance explained and the root criteria. In addition, they loaded on the factor with other items in the one-factor solution. Therefore, the one-factor model was the best model to be accepted. MC and CR items seemed to measure the same mathematical proficiency.

According to the fit statistics in Table 4-4, the reduction of the G^2 statistics seemed to be inflated because of the large sample size (N = 2,011). Therefore, the G^2 statistics was not used to determine the dimensionality. Based on the factor loadings, the variance explained, and the largest root value, it seemed that hypothesis one was supported.

Table 4-4

Change of the Likelihood Ratio G^2 in the Item Factor Analysis

Factor	G^2	Df	p
2 vs. 1	180.16	28	0.000
3 vs. 2	178.00	27	0.000

Confirmatory Approach

Two models were tested to address the first research question. The two-dimension format model was compared with the unidimensional model using MRCMLM. Chi-square difference test was used to compare the models based on the deviance index because the two models were hierarchical. According to the deviance indices in Table 4-5, it seemed that the unidimensional model fit the data much better than the two-dimension format model, $\chi^2_{(3)} = 2677.57$, $p < 0.001$. The correlation between the two dimensions was 0.89, indicating that the two dimensions measured similar constructs.

Table 4-5

Comparison of Unidimensional and Two-dimension Format Models Based on MRCMLM[a] (Data One)

Item	CL[b]	SC[c]	Estimate[d]	Error	One-dimension		Two-dimension	
					MNSQ[e]	Wfit[f]	MNSQ	Wfit
1. MC1	K	FN	-1.281	0.055	0.99	-0.2	**1.42**	**22.2**[g]
2. MC2	C	FN	-0.012	0.049	1.08	**4.7**	1.04	2.8
3. MC3	K	WN	-0.055	0.049	0.98	-1.1	0.95	-3.2
4. MC4	K	WN	-1.872	0.063	0.98	-0.6	0.93	-1.9
5. MC5	K	ME	-1.302	0.055	1.09	2.8	1.03	0.9
6. MC6	C	ME	0.398	0.049	1.04	2.6	1.02	1.2
7. MC7	K	WN	-1.399	0.057	1.02	0.7	0.99	-0.2
8. MC8	K	FN	0.178	0.049	1.02	1.2	0.98	-1.3
9. MC9	H	WN	-0.303	0.049	**1.18**	**9.4**	1.13	**7.2**
10. MC10	K	WN	-0.216	0.049	0.99	-0.5	0.96	-2.5
11. MC11	K	ME	0.229	0.049	1.08	**4.5**	1.05	3.0
12. MC12	C	FN	0.899	0.052	1.04	1.8	1.04	2.1
13. MC13	C	ME	-0.009	0.049	1.13	**7.1**	1.05	3.4
14. MC14	C	WN	-1.836	0.063	0.96	-1.1	0.91	-2.5
15. MC15	H	ME	0.274	0.049	1.11	**6.4**	1.06	**4.0**
16. MC16	C	WN	0.827	0.051	0.95	-2.8	0.94	-2.9
17. MC17	H	ME	1.992	0.066	1.03	0.6	1.01	0.3
18. CR1	H	ME	0.012	0.049	1.03	1.9	1.07	3.8
19. CR2	K	FN	-0.209	0.049	1.07	3.7	1.13	**6.3**
20. CR3	H	FN	-0.009	0.049	0.85	**-9.4**	0.86	**-7.6**
21. CR4	H	FN	0.856	0.052	0.86	**-7.1**	**1.96**	**27.6**
22. CR5	H	FN	-1.153	0.054	0.99	-0.5	1.06	1.9
23. CR6	K	WN	-0.937	0.052	0.96	-1.5	0.99	-0.5
24. CR7	H	FN	0.188	0.049	1.05	3.2	1.12	**5.8**
25. CR8	K	WN	0.593	0.050	0.85	**-8.7**	0.86	**-7.0**
26. CR9	K	WN	-0.183	0.049	0.93	**-4.1**	0.99	-0.5
27. CR10	H	WN	0.880	0.052	1.06	2.8	1.10	**4.3**
28. CR11	H	WN	0.407	0.049	0.90	**-5.8**	0.92	**-4.3**
29. CR12	K	ME	0.883	0.052	0.92	**-4.1**	0.97	-1.2
Deviance						68282.62		70960.19
Estimated[h] parameters						31		34
Correlation (dimension)						-		0.89

Note. [a] 29 items from Booklet 5, Gr.3 and Gr.4, TIMSS (N = 2,011). [b] CL-cognitive level, K-knowledge, C-complex procedure, H-higher mental process. [c] SC-sub-content, FN-fraction and number sense, ME-measurement, WN-whole number. [d] Estimate-difficulty parameter, [e] MNSQ-mean square residual. [f]Wfit-weighted fit. [g]Bold indicates misfit items. [h] relative magnitudes discussed in chapter 5.

The weighted fit indices showed that the unidimensional and two-dimension models did not fit the data very well. For the unidimensional model, five MC items (#2, #9, #11, #13, and #15) and 6 CR items (#3, #4, #8, #9, #11, and #12) were serious misfit items based on the criteria that any item is serious misfit item if the fit index is either larger than 4.00 or smaller than –4.00. For the two-dimension format model, three MC items (#1, #9, and #15) and 7 CR items (#2, #3, #4, #7, #8, #10, and #11) were serious misfit items. The reasons for the misfit vary. The assumptions of equal discrimination and zero guessing might be violated for MC items. In addition, the misfit index might be inflated by the large sample size.

Based on the latent distribution of the two-dimension format model (Figure 5), the latent distribution of the first dimension (MC items) were more negatively skewed than the distribution of the second dimension (CR items). The students' distribution for the second dimension showed greater spread than the distribution of the first dimension.

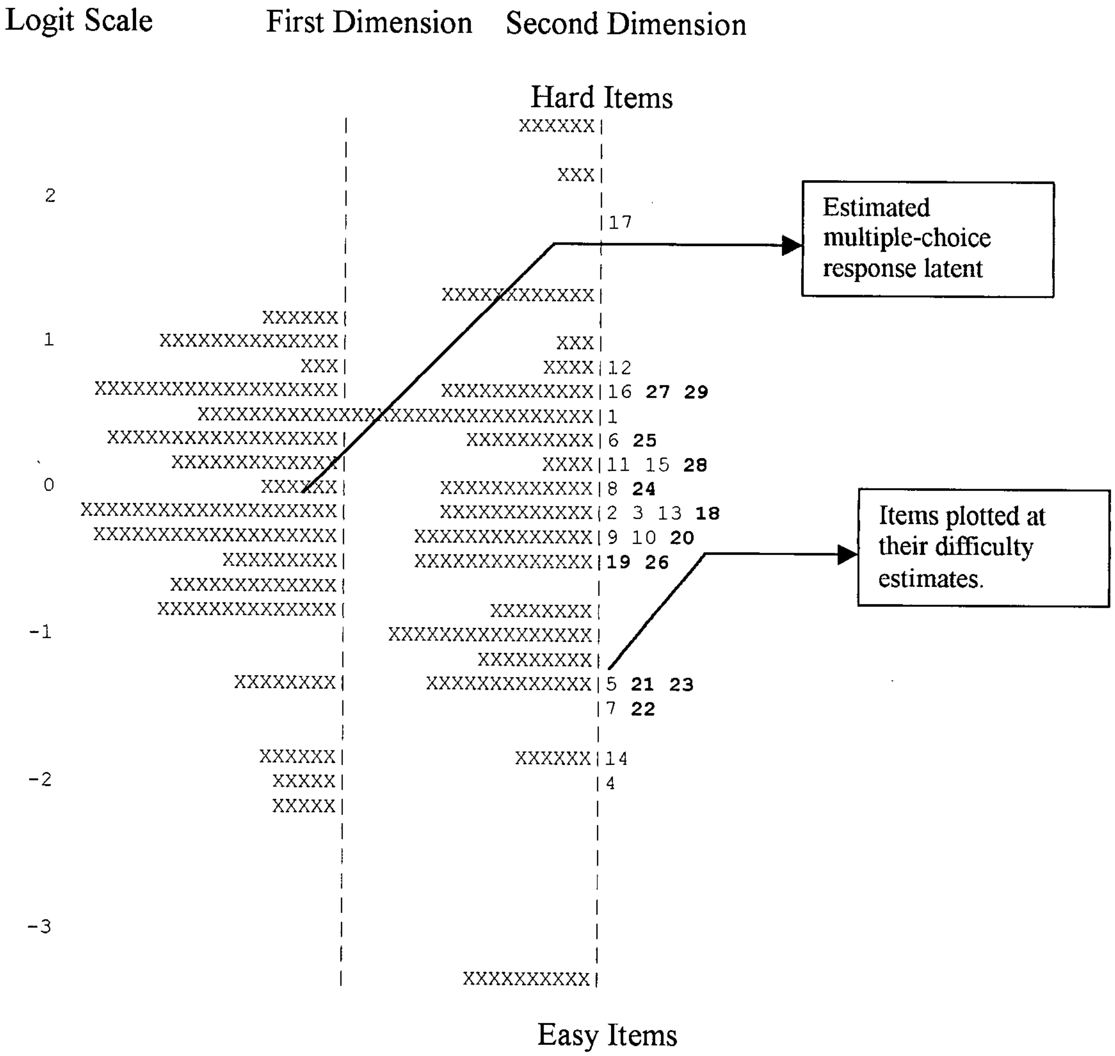


Figure 5. Map of Latent Distribution and Response Model Parameter Estimates for Two-dimension Format Model (N = 2,011).
Bold items are constructed-response items, non-bold items are multiple-choice items (29 items from TIMSS, Gr.3 and Gr.4 Mathematics Examination).

Summary for Research Question One

According to the evidence from both FIFA and MRCMLM models, it seemed that the data structure was unidimensional. The results from FIFA (exploratory factor analysis) indicated that the data structure could be explained by one dominant factor—mathematical proficiency. The examination of the factor loadings revealed: (1) Item local dependence was the main reason for the two non-significant minor dimensions; (2) the four items that failed to load on the dominant factor with other items were those that required either little or no computation skills. The results from MRCMLM also indicated that the unidimensional model fit much better than the two-dimension format model. The hypothesis that MC and CR items represented two dimensions (hypothesis two) was not supported.

RESEARCH QUESTION TWO

Exploratory Factor Analysis

Three hypotheses were tested to investigate whether MC and CR items differ in measuring students' cognitive ability beyond knowledge level in mathematics.

Hypothesis One: The test structure is unidimensional.

Hypothesis Two: The test structure is two-dimensional.

Hypothesis Three: The test structure is three-dimensional.

FIFA was applied as an exploratory approach. Sixteen items designed to tap high cognitive levels (complex procedures and high mental process) in TIMSS Mathematics examination (Gr.3 and Gr.4) were selected for the investigation. According to the variance and root criteria in Table 4-6, it seemed that the one-factor model was the best of all the three models. The total variance accounted for by the one-factor model was 24.8%. The largest root was 3.93, which was larger than 1.4, indicating a significant factor. Hypothesis one was supported. The first factor in the two-factor model was significant ($\lambda = 3.97$), but the second factor was not ($\lambda = 0.57$). Therefore, hypothesis two was rejected. Similarly, the first factor of the three-factor model was significant ($\lambda = 4.02$), but the second ($\lambda = 0.96$) and third factors ($\lambda = 0.58$) of the three-factor model were not significant. Hypothesis three was rejected.

Table 4-6

Comparison of Factor Loadings of the Three Solutions Based on FIFA[a] (Data One)

			One-factor[d]	Two-factor[e]		Three -factor		
Item	CL[b]	SC[c]	Factor 1	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3
1. MC2	C	FN	**0.393**	0.262	0.179	**0.395**	-0.005	0.039
2. MC6	C	ME	**0.396**	-0.009	**0.522**	**0.408**	-0.021	0.028
3. MC9	H	FN	0.243	0.103	0.182	0.212	-0.021	0.073
4. MC12	C	FN	**0.352**	0.140	0.273	**0.325**	0.016	0.052
5. MC13	C	ME	0.279	0.195	0.115	-0.098	**1.023**	0.000
6. MC14	C	WN	**0.533**	**0.422**	0.160	**0.425**	0.057	0.148
7. MC15	H	ME	0.243	0.082	0.208	0.138	0.046	0.124
8. MC16	C	WN	**0.544**	0.140	**0.526**	**0.615**	-0.002	-0.040
9. MC17	H	ME	**0.390**	-0.140	**0.664**	**0.620**	0.065	-0.292
10. CR1	H	ME	**0.475**	**0.322**	0.209	**0.377**	-0.039	0.185
11. CR3	H	FN	**0.824**	**0.734**	0.146	**0.776**	-0.044	0.150
12. CR4	H	FN	**0.844**	**0.721**	0.187	**0.776**	-0.044	0.150
13. CR5	H	FN	**0.568**	**0.677**	-0.109	**0.358**	0.044	0.291
14. CR7	H	FN	**0.393**	0.287	0.146	0.230	0.043	0.219
15. CR10	H	WN	**0.365**	**0.492**	-0.130	-0.109	0.007	**0.658**
16. CR11	H	WN	**0.613**	**0.625**	0.021	0.172	-0.002	**0.639**
Variance			24.81%	25.07%	3.44%	26.25%	5.96%	3.53%
Largest roots			3.93	3.97	0.57	4.02	0.96	0.58
Correlation								
Factor 1				1		1		
Factor 2			-	0.609	1	0.303	1	
Factor 3						0.566	0.204	1

Note. [a] 16 items from Booklet 5, Gr.3 and Gr.4, TIMSS (N = 2,011). [b] CL-cognitive level, C-complex procedure, H-higher mental process. [c] SC-sub-content, FN-fraction a number sense, ME-measurement, WN-whole number. [d] Factor loadings for one-factor model are principal factor loadings. [e] Factor loadings for two- and three-factor models are promax factor loadings. [f] Bold numbers are loadings greater than the absolute value of 0.30.

Examination of Factor Loadings

All the loadings of the one-factor solution were above 0.30 except for three MC items (#9, #13 and #15, see p.51). The examination of these three items revealed that they required no calculation, indicating that students can answer these three questions based on their common sense and knowledge.

According to the three-factor solution, most MC and all CR items loaded on the first factor and the non-significant minor dimensions were mainly represented by 1 MC #13 and 2 CR items (#10 and #11). CR #10 and CR #11 were two locally dependent items based on the high loadings (both loaded highly on factor 2) and content analysis. CR#11 required students to provide the explanation of the answer to CR10. The examination of MC #13 revealed that this item required no computation skill at all because the examinees can answer it correctly based on their common sense knowledge.

Because the two minor dimensions were not significant based on the variance and root criteria, the one-factor model was the best model to be accepted. Thus, MC and CR beyond knowledge level seemed to measure same mathematical proficiency.

Confirmatory Approach

Three models were tested using MRCMLM for the second research question:

Hypothesis One (Rasch Model): The test structure is unidimensional.

Hypothesis Two (Between-item Model): The test structure is two-dimensional
(MC vs. CR related proficiency).

Hypothesis Three (Within-item CR Model): The test structure is two-dimensional
(Overall Mathematics proficiency vs. CR proficiency).

The two-dimension between-item (format) model was compared with the unidimensional model by looking at the difference between the deviance indices of the two models. The two-dimension model was significantly better than the unidimensional model, $\chi^2_{(3)} = 249.47$, $p < 0.001$. Such result seemed to contradict the result of FIFA. However, it was possible that the deviance difference between the two models might be inflated by the large sample size (N = 2,011). The correlation between the two dimensions (MC vs. CR) was 0.89, indicating that the two dimensions measured similar constructs.

As shown in Table 4-7, the two-dimension within CR model seemed to be better than the other two models because the deviance was 19.76 smaller than that of the two-dimension between-item model, indicating that CR items might have some additional characteristics that differed from the overall mathematics proficiency. However, the correlation between the two dimensions (0.84) seemed to indicate that the two dimensions measured similar ability. Thus, the results should be interpreted with caution because the large sample size might inflate the deviance difference (19.76) between the two models.

Table 4-7

Comparison of Unidimensional and Two-dimension Format Models Based on MRCMLM[a] (Data One)

			One-dimension	Two-dimension	Two-dimension
			MNSQ[d]	Between-item	Within-item
Item	CL[b]	SC[c]	(Wfit[e])	MNSQ (Wfit)	MNSQ (Wfit)
1. MC2	C	FN	1.03 (2.0)	0.98 (-1.2)	1.01 (0.6)
2. MC6	C	ME	1.01 (0.7)	0.98 (-1.2)	1.00 (0.3)
3. MC9	H	FN	1.09 **(4.9)**[6]	1.03 (2.0)	1.03 (2.0)
4. MC12	C	FN	1.02 (0.8)	1.01 (0.5)	1.00 (-0.2)
5. MC13	C	ME	1.07 **(4.1)**	1.02 (1.8)	1.02 (1.5)
6. MC14	C	WN	1.01 (-0.1)	0.93 (-2.0)	0.93 (-1.9)
7. MC15	H	ME	1.09 **(5.6)**	1.04 **(3.2)**	1.05 **(3.8)**
8. MC16	C	WN	0.95 **(-2.8)**	0.97 (-1.3)	0.96 **(-2.4)**
9. MC17	H	ME	1.01 (0.3)	1.05 (1.3)	1.01 (0.3)
10. CR1	H	ME	0.95 **(-2.8)**	1.00 (0.0)	1.02 (0.7)
11. CR3	H	FN	0.89 **(-7.2)**	0.90 **(-5.2)**	0.91 **(-4.8)**
12. CR4	H	FN	0.88 **(-6.5)**	0.90 **(-5.2)**	0.87 **(-6.0)**
13. CR5	H	FN	0.99 (-0.2)	1.05 (1.5)	1.08 **(2.4)**
14. CR7	H	FN	0.99 (-0.3)	1.00 (0.1)	1.02 (0.5)
15. CR10	H	WN	1.01 (0.5)	1.09 **(3.7)**	1.10 **(4.2)**
16. CR11	H	WN	0.92 **(-5.2)**	0.95 **(-2.8)**	0.95 **(-2.3)**
Deviance			37574.33	37324.86	37305.10
Estimated parameters			18	21	21
Correlation (dimension)			-	0.885	0.843

Note. [a] 16 items from Booklet 5, Gr.3 and Gr.4, TIMSS (N = 2,011). [b] CL-Cognitive Level, C-complex procedure, H-higher mental process. [c] SC-Sub-content, FN-fraction and number sense, ME-measurement, WN-whole number. [d] MNSQ-mean square residual. [e] Wfit-weighted fit. [f] Bold items are misfit.

Summary for Research Question Two

The results from the exploratory factor analysis (FIFA) supported the hypothesis that the test structure was unidimensional. MC and CR items seemed to measure same mathematical proficiency beyond knowledge level. The confirmatory factor analysis (MRCMLM) revealed that the test structure was two-dimensional (MC vs. CR). In addition, it seemed that CR items might tap distinct characteristics that differed from the overall mathematical proficiency. However, the results from the confirmatory factor analysis needed to be interpreted with caution because the deviance difference might be inflated due to the large sample size. The examination of the factor loadings revealed that the item local dependence and cognitive demand instead of format factor were the main reasons for the existence of the minor dimensions. The high correlation between the two dimensions (MC and CR items) also indicated that the two dimensions were similar in measuring students' ability. Therefore, the hypothesis that the test structure was unidimensional was supported.

RESEARCH QUESTION THREE

To test whether high ability students differ from low ability students in dealing with different formats, the confirmatory approach instead of the exploratory approach was used because it is more meaningful to confirm what previous researchers suggested that high and low ability students differed in dealing with test items. The three hypotheses tested are as follows:

Hypothesis One: MC and CR items are two dimensions for low ability students.

Hypothesis Two: MC and CR items are one dimension for high ability students.

Hypothesis Three: There is statistically significant interaction between format and ability level.

Hypothesis One

As presented in Table 4-8, the deviance of the two-dimension format model was smaller than that of the unidimensional model, indicating that the two-dimension model represented by MC and CR items fit the data better than the unidimensional model for low ability students, $\chi^2_{(3)} = 19.67$, $p < 0.01$. Hypothesis one was supported.

Table 4-8

Comparison of Unidimensional and Two-dimension Format Model on Low Ability Students' Responses[a] (Data One)

					One-dimension		Two-dimension	
Item[b]	CL[c]	SC[d]	Estimate[e]	Error	MNSQ[f]	Wfit[g]	MNSQ	Wfit
1.MC2	C	FN	0.926	0.091	1.01	0.2	1.01	0.4
2.MC6	C	ME	1.233	0.097	1.03	0.6	1.03	0.6
3.MC9	H	FN	0.293	0.083	1.03	1.9	0.97	-1.4
4.MC12	C	FN	1.622	0.109	1.02	0.3	1.02	0.3
5.MC13	C	ME	0.705	0.087	0.99	-0.3	1.00	0.1
6.MC14	C	WN	-0.465	0.085	1.00	0.0	1.01	0.3
7.MC15	H	ME	0.877	0.090	1.03	0.8	1.03	0.7
8.MC16	C	WN	2.175	0.132	1.04	0.4	1.02	0.2
9.MC17	H	ME	2.658	0.161	1.02	0.2	1.02	0.2
10.CR1	H	ME	1.114	0.094	0.99	-0.2	0.97	-0.6
11.CR3	H	FN	2.107	0.129	1.01	0.1	0.98	-0.2
12.CR4	H	FN	3.496	0.234	1.01	0.1	1.02	0.2
13.CR5	H	FN	0.067	0.083	0.99	-0.9	1.00	0.0
14.CR7	H	FN	1.105	0.094	1.02	0.3	1.02	0.3
15.CR10	H	WN	1.936	0.121	1.01	0.2	1.00	0.1
16.CR11	H	WN	2.158	0.131	1.02	0.3	0.99	-0.1
Deviance						9544.55		9524.88
Estimated Parameters						18		21
Correlation (dimension)						-		-0.070

Note. [a] Low ability - students at the low 30% of the total score (N = 651); [b] 16 items from Booklet 5, Gr.3 and Gr.4, TIMSS. [c] CL-cognitive level, C-complex procedure, H-higher mental process. [d] SC-sub-content, FN-fraction and number sense, ME-measurement, WN-whole number. [e] Estimate-difficulty parameter, [f]MNSQ-mean square residual. (7) Wfit-weighted fit.

Hypothesis two

As shown by the deviance indices in Table 4-9, the two-dimension format model and the unidimensional model seemed to fit the data similarly for the high ability students ($\chi^2_{(3)} = 10.72$, $p > 0.05$). The weighted fit indices of 16 items were within normal range. Hypothesis two was supported.

Table 4-9

Comparison of Unidimensional and Two-dimension Format Models on High Ability Students' Responses[a] (Data One)

					One-dimension		Two-dimension	
Item[b]	CL[c]	SC[d]	Estimate[e]	Error	MNSQ[f]	Wfit[g]	MNSQ	Wfit
1. MC2	C	FN	-0.962	0.088	1.01	0.3	1.01	0.3
2. MC6	C	ME	-0.705	0.084	0.99	-0.4	0.99	-0.5
3. MC9	H	FN	-0.849	0.086	1.01	0.3	1.01	0.4
4. MC12	C	FN	-0.044	0.079	1.02	1.5	1.02	1.5
5. MC13	C	ME	-0.691	0.084	1.03	0.9	1.00	0.1
6. MC14	C	WN	-3.005	0.179	0.97	-0.1	0.98	-0.0
7. MC15	H	ME	-0.441	0.081	1.03	1.6	1.00	0.1
8. MC16	C	WN	-0.534	0.082	1.03	1.0	1.00	-0.1
9. MC17	H	ME	0.903	0.087	1.00	-0.1	0.98	-0.4
10. CR1	H	ME	-1.114	0.091	1.00	-0.0	1.01	0.2
11. CR3	H	FN	-1.907	0.115	0.99	-0.1	1.01	0.2
12. CR4	H	FN	-0.849	0.086	0.99	-0.3	0.97	-0.8
13. CR5	H	FN	-2.537	0.147	0.99	-0.1	1.02	0.2
14. CR7	H	FN	-0.726	0.084	1.00	0.0	1.01	0.3
15. CR10	H	WN	-0.057	0.079	1.04	2.4	1.04	2.4
16. CR11	H	WN	-1.009	0.089	0.96	-1.1	1.02	0.4
Deviance						11928.78		11918.06
Estimated Parameters						18		21
Correlation (dimension)						-		0.221

Note. [a] High ability - students at the high 30% of the total score (N = 609); [b] 16 items from Booklet 5, Gr.3 and Gr.4, TIMSS. [c] CL-cognitive level, C-complex procedure, H-higher mental process.
[d] SC-sub- content, FN-fraction and number sense, ME-measurement, WN-whole number.
[e] Estimate-difficulty parameter, [f] MNSQ-mean square residual. [g] Wfit-weighted fit

Hypothesis Three

Descriptive statistics as well as ANOVA analysis were used to see how high ability students differed from low ability students in dealing with different formats. As shown in Table 4-10, low ability students' item mean scores of MC and CR items were lower than that of high ability students. Low ability students' item mean score of MC items was higher than that of CR items, whereas high ability students' item mean score of CR items was higher than that of MC items.

Table 4-10

Mean Scores of High and Low Ability Students' Responses on MC and CR Items (Data One)

Format	Ability[a]	Item Mean	Std. Deviation	N
Multiple-choice	low	0.2869	0.1403	651
	high	0.6460	0.1593	609
	Total	0.4605	0.2338	1260
Constructed-response	low	0.2065	0.1590	651
	high	0.7445	0.1705	609
	Total	0.4666	0.3153	1260

Note. [a] Low ability - students at the low 30% of the total score; High ability -students at the high 30% of the total score.

In Table 4-11, the test statistics showed that high and low ability students differed on their performances on different formats. The interaction of ability group and format was statistically significant, $F_{(1, 1258)} = 208.20$, $p < 0.001$, indicating that high ability students differed from low ability students in dealing with MC and CR items (see Figure 6). The effect size f for the interaction was regarded as large (Cohen, 1988, p.287), so hypothesis three was supported.

Table 4-11

Analysis of Variance for High and Low Ability Students' Mean Scores on MC and CR Items (Data One)

Source	SS	df	MS	F	Eta (η^2)	f
Between			Between			
Group	126.639	1	126.639	4995.503*	0.799	1.993
Error (Group)	31.891	1258	0.025			
Within			Within			
Format[a]	0.051	1	0.050	2.130	0.002	0.000
Format * Group[b]	5.036	1	5.036	208.199*	0.142	0.412
Error (Format)	30.428	1258	0.024			

Note. [a] Format-MC and CR; [b] Low ability-students at the low 30% of the total score (N = 651); High ability-students at the high 30% of the total score (N = 609). * p < .001.

Summary for Research Question Three

The results from the confirmatory approaches indicated that, for low ability students, the two-dimension between-item format model was a better model. It appeared that the two formats represented two somewhat different constructs for the low ability students. However, for high ability students, the two-dimension format model was no better than the unidimensional model. Hypothesis one and hypothesis two were supported. Based on the ANOVA analysis, there was a statistically significant interaction found between ability and format, indicating that the two groups did differ in dealing with different formats (MC vs. CR). Hypothesis three was supported as well.

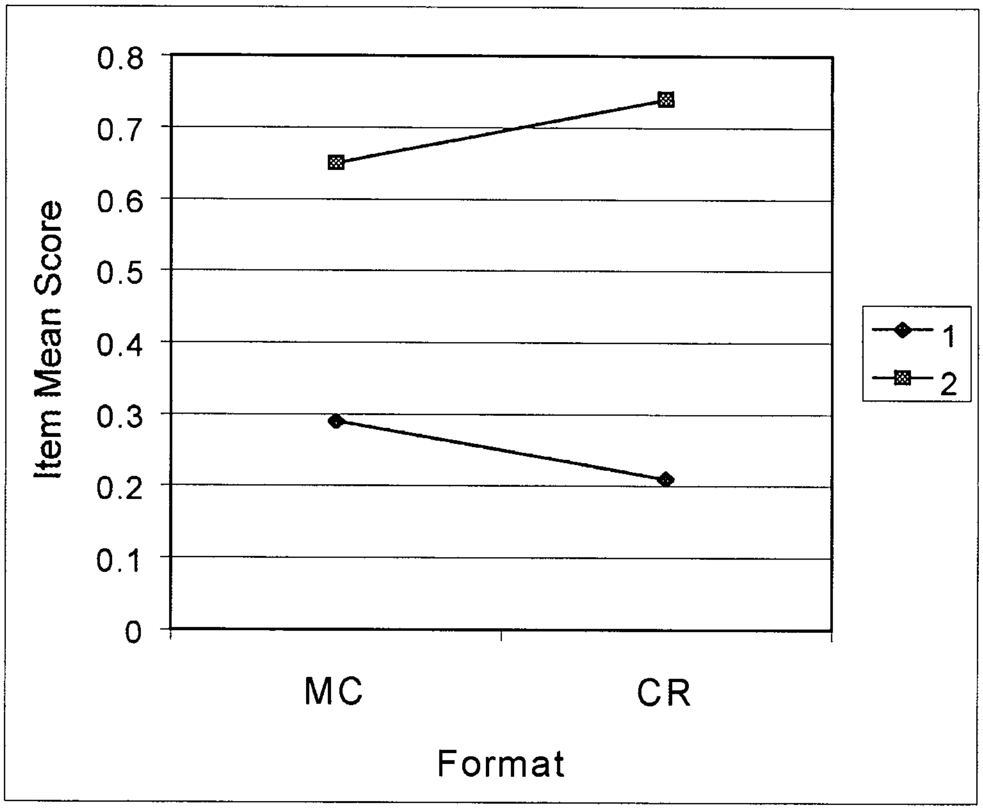


Figure 6. Comparison of High and Low Ability Students in terms of Format Differences (Data Set One). MC – Multiple-Choice, CR – Constructed-Response, 1 – Low ability, 2 – High ability

EXAMINATION OF DATA SET TWO: TIMSS (GRADE 7 AND GRADE 8)

Missing Data

The second analyses were based on the TIMSS data from Booklet 3, Grade 7 and Grade 8. Twenty-six items were selected for the investigation. The Canadian sample (N = 2,073) available from the TIMSS data bank was used for the investigation. By looking at the missing rate in Table 4-12, it can be seen that students omitted more CR items than MC items. The average response rates for MC and CR items were 98.9% and 88.8%, respectively. The CR #8 had the highest missing rate (16.6%) and not-attempted response rates (9.1%). Speededness did not seem to be a problem because no item had more than 10% of not-attempted responses. One CR #8 had more than 10% of missing response. Therefore, all the missing data and not-attempted responses were treated as wrong responses and coded as 0 in the data file.

Constructed-Response Items

Three CR items, coded as 0, 1, and 2, were dichotomized. For the two-point open-ended response items, scores of zero, one were re-coded to a value of zero, while score of two was re-coded to a value of one. The information loss, examined using FLUX software, was only 3% (see Figure 7, Appendix B). Therefore, it was concluded that dichotomizing 3 CR items in the study did not lead to the loss of information.

Table 4-12

Percentage of Missing Data (Data Set Two)

Item	Valid Response Rate	Percent of Missing Response	Percent of not Attempted
1. MC1	99.5	0.5	0.0
2. MC2	99.7	0.3	0.0
3. MC3	99.5	0.5	0.0
4. MC4	99.8	0.2	0.0
5. MC5	98.2	1.8	0.0
6. MC6	99.6	0.4	0.0
7. MC7	99.8	0.2	0.0
8. MC8	99.0	1.0	0.0
9. MC9	97.7	2.3	0.0
10. MC10	97.2	2.8	0.0
11. MC11	98.7	1.3	0.0
12. MC12	99.8	0.2	0.0
13. MC13	98.6	1.4	0.0
14. MC14	98.3	1.7	0.0
15. MC15	98.9	1.1	0.0
16. MC16	98.5	1.5	0.0
17. MC17	99.0	1.0	0.0
18. MC18	99.1	1.0	0.0
Mean	98.9	1.1	0.0
19. CR1	93.3	6.7	0.0
20. CR2	92.5	7.5	0.0
21. CR3	89.7	9.3	1.1
22. CR4	87.3	8.1	4.7
23. CR5	95.2	3.7	1.1
24. CR6	87.9	9.4	2.7
25. CR7	90.1	5.9	4.0
26. CR8	74.3	**16.6**	**9.1**
Mean	88.8	8.4	2.8

Note. Items were selected from TIMSS, Gr.7 and Gr.8 Mathematics Examination (Booklet 3). Sample size is 2,073.

Results from Classical Test Theory

From Table 4-13, it can be seen that students scored higher on MC items than on CR items (the item mean score for MC items was 0.58; the item mean score for CR items was 0.38). The item-total correlation varied from 0.16 to 0.55. On average, CR items had higher item total correlations than MC items. The overall reliability (Cronbach's alpha) was 0.85.

Table 4-13

Descriptive Item Analysis[a] (Data Set Two)

Item	CL[b]	SC[c]	Item Mean	Std Dev	Item Total Correlation[d]
1. MC1	K	FN	0.60	0.49	0.53
2. MC2	H	AB	0.73	0.44	0.31
3. MC3	C	ME	0.65	0.48	0.42
4. MC4	C	FN	0.74	0.44	0.48
5. MC5	K	AB	0.46	0.50	0.16
6. MC6	K	ME	0.82	0.39	0.31
7. MC7	K	FN	0.76	0.43	0.26
8. MC8	H	FN	0.57	0.50	0.37
9. MC9	H	FN	0.70	0.46	0.47
10. MC10	K	AB	0.56	0.50	0.39
11. MC11	K	ME	0.38	0.49	0.40
12. MC12	K	ME	0.74	0.44	0.32
13. MC13	H	FN	0.68	0.47	0.40
14. MC14	K	AB	0.27	0.45	0.30
15. MC15	C	FN	0.68	0.47	0.27
16. MC16	K	AB	0.40	0.49	0.17
17. MC17	C	FN	0.39	0.49	0.36
18. MC18	K	FN	0.39	0.49	0.51
19. CR1	H	FN	0.71	0.45	0.37
20. CR2	H	ME	0.33	0.47	0.55
21. CR3	H	AB	0.30	0.89	0.51
22. CR4	H	AB	0.27	0.44	0.53
23. CR5	H	FN	0.44	0.50	0.41
24. CR6	H	FN	0.39	0.49	0.41
25. CR7	H	ME	0.30	0.82	0.51
26. CR8	H	ME	0.29	0.70	0.50
Mean for MC			0.58		
Reliability[5]			0.77		
Mean for CR			0.38		
Reliability			0.76		
Mean			0.52		
Reliability for MC and CR			0.85		

Note. [a] 26 items from Booklet 3, Gr.7 and Gr.8, TIMSS (N = 2,073). [b] CL-cognitive Level, K-knowledge, C-complex procedure, H-higher mental process. [c] SC-sub-content, FN-fraction and number sense, ME-measurement, AB-algebra. [d] Discrimination index: point-biserial. [e] Reliability-Cronbach's alpha.

RESEARCH QUESTION ONE

Exploratory Factor Analysis

Three models were tested to investigate whether the format differences affect the structure of the mathematics test.

Hypothesis one: The test structure is unidimensional.

Hypothesis two: The test structure is two-dimensional.

Hypothesis three: The test structure is three-dimensional.

FIFA was applied as an exploratory approach to test the hypotheses. By looking at the variance and largest root indices in Table 4-14, it seemed that the one-factor model was the best of all the three models. The total variance accounted for by the model was 32.9%, and the largest root was 8.56. Hypothesis one was supported. In terms of the two-factor model, the first factor was a significant factor ($\lambda = 8.60$), but the second factor was not ($\lambda = 1.08$). Hypothesis two was rejected. Similarly, the first factor was a significant factor ($\lambda = 8.66$), and the second ($\lambda = 1.13$) and third factors ($\lambda = 0.85$) in the three-factor solution were not significant factors. Therefore, hypothesis three was rejected.

Table 4-14

Comparison of Factor Loadings of the Three Solutions Based on FIFA (Data Two)

			One-factor[d]	Two-factor[e]		Three-factor		
Item[a]	CL[b]	SC[c]	Factor 1	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3
1. MC1	K	FN	**0.715**	**0.443**[f]	0.297	**0.651**	0.062	-0.051
2. MC2	H	AB	**0.440**	**0.397**	0.054	**0.446**	-0.045	-0.061
3. MC3	C	ME	**0.588**	**0.440**	0.166	**0.553**	-0.020	-0.058
4. MC4	C	FN	**0.713**	**0.550**	0.185	**0.628**	0.007	0.033
5. MC5	K	AB	0.202	0.193	0.016	0.171	-0.047	0.056
6. MC6	K	ME	**0.497**	**0.447**	0.065	**0.459**	-0.058	0.004
7. MC7	K	FN	**0.381**	0.279	0.117	**0.331**	0.006	-0.026
8. MC8	H	FN	**0.496**	**0.316**	0.196	**0.438**	0.053	-0.042
9. MC9	H	FN	**0.659**	**0.420**	0.266	**0.564**	0.034	-0.008
10. MC10	K	AB	**0.529**	**0.423**	0.127	**0.409**	0.047	0.077
11. MC11	K	ME	**0.536**	**0.322**	0.231	**0.555**	-0.030	-0.048
12. MC12	K	ME	**0.536**	**0.338**	0.133	**0.396**	0.010	0.029
13. MC13	H	FN	**0.555**	**0.523**	0.050	**0.387**	0.071	0.142
14. MC14	K	AB	**0.419**	0.231	0.204	**0.351**	0.024	0.056
15. MC15	C	FN	**0.383**	**0.307**	0.085	**0.364**	-0.030	-0.03
16. MC16	K	AB	0.232	0.160	0.083	0.167	0.005	0.025
17. MC17	C	FN	**0.497**	**0.398**	0.119	**0.365**	0.052	0.078
18. MC18	K	FN	**0.696**	**0.534**	0.181	**0.644**	-0.046	0.007
19. CR1	H	FN	**0.527**	**0.453**	0.094	**0.466**	-0.031	0.003
20. CR2	H	ME	**0.753**	**0.419**	**0.353**	**0.754**	-0.006	-0.082
21. CR3	H	AB	**0.779**	**-0.506**	**1.366**	-0.016	**1.005**	-0.017
22. CR4	H	AB	**0.817**	**-0.450**	**1.334**	0.027	**0.984**	0.003
23. CR5	H	FN	**0.535**	**1.129**	**-0.557**	-0.055	0.004	**0.886**
24. CR6	H	FN	**0.537**	**1.196**	**-0.620**	-0.054	-0.020	**0.916**
25. CR7	H	ME	**0.709**	**0.467**	0.264	**0.675**	-0.005	-0.005
26. CR8	H	ME	**0.706**	**0.557**	0.171	**0.657**	-0.032	0.060
Variance			32.92%	33.09%	4.39%	23.65%	7.09%	4.32%
Largest roots			8.56	8.60	1.08	8.66	1.13	0.85
Correlation								
Factor 1				1		1		
Factor 2				0.837	1	0.407	1	
Factor 3						0.318	0.386	1

Note. [a] 26 items from Booklet 3, Gr.7 and Gr.8, TIMSS (N = 2,073). [b] CL-cognitive level, K-knowledge, C-complex procedure, H-higher mental process. [c] SC-sub-content, FN-fraction and number sense, ME-measurement, AB-whole number. [d] Factor loadings for one-factor model is principal factor loadings. [e] Factor loadings for two- and three-factor models are promax factor loadings. [f] Bold numbers are loadings greater than the absolute value of 0.30.

Examinations of Factor Loadings

By looking at the loadings of the one-factor solution, all the items loaded on the factor had loadings greater than 0.30 except for two MC items (#5 and #16). The examination of these 2 items showed that they had medium item difficulty (0.46 and 0.40) and low item discrimination (0.16 and 0.17).

MC #5

The cost, C, of printing greeting cards consists of a fixed charge of 100 cents and a charge of 6 cents for each card printed. Which of these equations can be used to determine the cost of printing *n* cards?

A. $C = (100+6n)$ cents B. $C = (106 + n)$ cents C. $C = (6 + 100n)$ cents D. $C = (106n)$ cents
E. $C = (600n)$ cents

MC #16 $\frac{x}{2} < 7$ is equivalent to

A. $x < \frac{7}{2}$ B. $x < 5$ C. $x < 14$ D. $x > 5$ E. $x > 14$

The examination of the 2 MC items revealed that both of them were algebra problems. In terms of MC #5, the examinees needed to translate the words of the problem into an internal representation (e.g., equation). In terms of MC #16, knowledge of the arithmetic rule was required to produce the correct answer. For both items, the examinees did not need to demonstrate computation skills. By looking at other items that had higher loadings (e.g., CR #3 and CR #4), it seemed either high level of computation skills or multiple computation processes were involved.

The exploratory two- and three-factor solution revealed that the minor dimensions were not significant based on the variance and root criteria. However, two pairs of CR items (#3 and #4, #5 and #6) were found to be locally dependent.

CR #3 and CR #4.

There are *54* kilograms of apples in two boxes. The second box of apples weights *12* kilograms more than the first. How many kilograms of apples are in each box? *Show* your work.

CR Item #5 and #6.

Teresa wants to record *5* songs on tape. Estimate the nearest minute the total time taken for all five songs to play and *explain* how this estimate was made.

Song	Amount of Time
1	2 minutes 41 seconds
2	3 minutes 10 seconds
3	2 minutes 51 seconds
4	3 minutes
5	3 minutes 32 seconds

CR Items (#3 and #4) were complex problem solving tasks. First, students were required to use schema knowledge to represent the structure of the problem. Second, schema knowledge (rule) of the mathematical equation of the variables that matched the problem structure was needed. Third, computation skill was required to produce the correct answer. The two items were locally dependent due to the same problem context. The success of one item meant the success of the other. Similarly, CR items #5 and #6 were two locally dependent items due to the same context.

The two pairs of items loaded highly on the factor in the one-factor solution. In addition, the two minor dimensions represented by the two pairs of items were not significant according to the variance and the root criteria. Therefore, the one-factor model was the best model to be accepted. MC and CR items seemed to measure similar mathematical proficiency.

Although the reduction of the G^2 statistics seemed to suggest that the two-factor model provided a better fit to the data than the one-factor model, and the three-factor model was better than the two-factor model (see Table 4-15). The G^2 index was not a good index because it was inflated by the large sample size (N = 2073).

Table 4-15

Change of the Likelihood Ratio G^2 in the Item Factor Analysis

Factor	G^2	df	p
2 vs. 1	883.69	25	0.000
3 vs. 2	760.17	24	0.000

Confirmatory Approach

MRCMLM was applied to compare the two-dimension format model with the unidimensional model. According to the deviance indices in Table 4-16, it seemed that the unidimensional model fit the data better than the two-dimension format model, $\chi^2_{(3)} = 2445.02$, $p < 0.001$.

Table 4-16

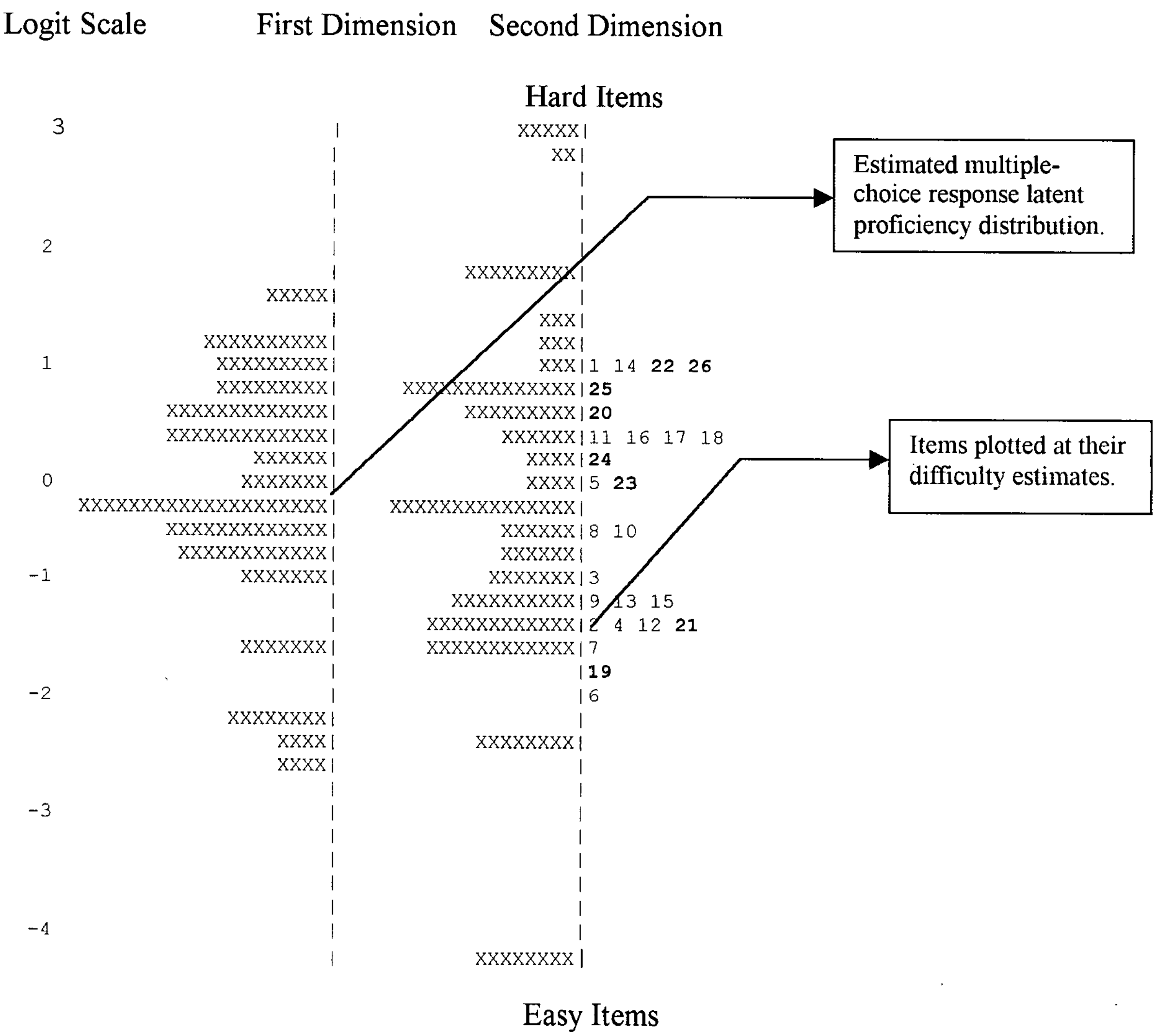
Comparison of Unidimensional and Two-dimension Format Models Based on MRCMLM (Data Two)

					Unidimensional		Two-dimension	
Item[a]	CL[b]	SC[c]	Estimate[d]	Error	MNSQ[e]	Wfit[f]	MNSQ	Wfit
1. MC1	K	FN	-0.492	0.050	0.90	**-4.9**[g]	**1.58**	**21.5**
2. MC2	H	AB	-1.234	0.054	1.07	2.4	1.04	1.4
3. MC3	C	ME	-0.763	0.051	0.98	-0.8	0.97	-1.5
4. MC4	C	FN	-1.294	0.055	0.88	**-4.2**	0.87	**-4.9**
5. MC5	K	AB	0.200	0.049	**1.25**	**12.6**	**1.22**	**11.7**
6. MC6	K	ME	-1.834	0.061	1.05	1.4	1.02	0.6
7. MC7	K	FN	-1.443	0.056	1.13	**4.0**	1.12	3.8
8. MC8	H	FN	-0.356	0.050	1.06	2.9	1.03	1.7
9. MC9	H	FN	-1.030	0.053	0.93	-2.7	0.91	-3.5
10. MC10	K	AB	-0.314	0.050	1.02	1.0	1.00	0.2
11. MC11	K	ME	0.612	0.051	1.00	0.0	0.98	-1.3
12. MC12	K	ME	-1.303	0.055	1.11	3.5	1.07	2.3
13. MC13	H	FN	-0.967	0.052	1.02	0.7	1.00	-0.1
14. MC14	K	AB	1.221	0.055	1.03	1.4	1.02	0.7
15. MC15	C	FN	-0.937	0.052	1.14	**5.3**	1.12	**4.7**
16. MC16	K	AB	0.502	0.050	**1.22**	**10.9**	**1.17**	**9.0**
17. MC17	C	FN	0.563	0.050	1.04	1.8	1.00	-0.2
18. MC18	K	FN	0.555	0.050	0.90	**-5.6**	0.88	**-6.6**
19. CR1	H	FN	-1.115	0.053	1.03	1.0	1.16	**5.1**
20. CR2	H	ME	0.891	0.052	0.85	**-7.3**	0.89	**-4.7**
21. CR3	H	AB	1.069	0.053	0.87	**-5.8**	**1.92**	**27.9**
22. CR4	H	AB	1.254	0.055	0.86	**-6.0**	0.85	**-5.6**
23. CR5	H	FN	0.310	0.050	1.00	-0.1	1.08	3.4
24. CR6	H	FN	0.566	0.050	0.98	-1.0	1.06	2.5
25. CR7	H	ME	1.044	0.053	0.89	**-5.1**	0.94	-2.2
26. CR8	H	ME	1.124	0.054	0.90	**-4.3**	0.90	-3.8
Deviance						60591.30		63036.32
Estimated[h] parameters						28		31
Correlation (dimension)						-		0.907

Note. [a] 26 items from Booklet 5, Gr.7 and Gr.8, TIMSS (N = 2,073). [b] CL-cognitive level, K-knowledge, C-complex procedure, H-higher mental process. [c] SC-sub-content, FN-fraction and number sense, ME-measurement, AB-whole number. [d] Estimate-difficulty parameter, [e] MNSQ-mean square residual. [f] Wfit-weighted fit. [g] Bold indicates misfit items. [h] the magnitudes are discussed in chapter 5.

The correlation between the two dimensions was 0.91, indicating that the two dimensions measured similar ability. According to the weighted fit indices, it seemed that both models did not fit the data very well. The unidimensional model seemed to be better than the two-dimension format model. However, the fit index of the unidimensional model showed that 12 out of 26 items were found to be misfit items. The high percentage of misfit items might be due to some other reasons (failure to vary guessing, discrimination parameters, or large sample size).

Based on the latent distribution of the two-dimension format model (Figure 9), the first dimension (MC items) was more negatively skewed than the second dimension (CR items). Students' distribution for the second dimension showed greater spread than the distribution of the first dimension.



Estimated multiple-choice response latent proficiency distribution.

Items plotted at their difficulty estimates.

Figure 9. Map of Latent Distribution and Response Model Parameter Estimates for Two-dimension Format Model. Bold items are constructed-response items, non-bold items are multiple-choice items (26 items from TIMSS, Gr.7 and Gr.8 Mathematics Examination).

Summary for Research Question One

The results from FIFA (exploratory factor analysis) indicated that the data structure was unidimensional. The results from MRCMLM also revealed that the unidimensional model was much better than the two-dimension format model. In addition, the examination of factor loadings showed that the existence of the non-significant minor dimensions were due to the local dependence of two pairs of CR items instead of format differences. The 2 MC items that failed to load on the dominant factor with other items were those that required no computation skills. The hypothesis that MC and CR items represented two distinct factors (hypothesis two) was not supported.

RESEARCH QUESTION TWO

Exploratory Factor Analysis

Three hypotheses were tested to investigate whether MC and CR items differ in measuring students' cognitive ability beyond knowledge level in mathematics.

Hypothesis One: The test structure is unidimensional.

Hypothesis Two: The test structure is two-dimensional.

Hypothesis Three: The test structure is three-dimensional.

Full-Information Factor Analysis was applied as an exploratory approach. Sixteen items designed to tap high cognitive levels (complex procedures and high mental process) in the TIMSS Mathematics Examination (Gr.7 and Gr.8) were selected for the investigation. By looking at the variance and largest root indices in Table 4-17, it seemed that the one-factor model was the best of all the three models. The total variance accounted for by the model was 38.3% and largest root was 6.03. In terms of the two-factor model, the first dimension was a significant factor ($\lambda = 6.10$), but the second dimension was not ($\lambda = 1.04$). Hypothesis two was rejected. Similarly, the first dimension of the three-factor model was significant ($\lambda = 6.18$), but the second ($\lambda = 1.12$) and third ($\lambda = 0.77$) factors were not. Hypothesis three was rejected.

Examination of Factor Loadings

All the loadings of the one-factor solution were above 0.30, indicating that these 16 items measured the same cognitive mathematical proficiency beyond knowledge level. The exploratory two- and three-factor solution revealed that the minor dimensions were not significant. The loadings in the two-factor solution seemed to indicate that 2 CR items loaded highly on the minor dimension. However, the second dimension in the two-factor solution was not significant. By looking at the exploratory three-factor solution, all MC and 4 CR items loaded on the first factor whereas the non-significant minor dimensions were represented by two pairs of locally dependent items (CR #3 and #4, CR #5 and #6). The minor dimensions represented by the two pairs of locally dependent items did not seem to be significant based on the variance and root criteria. In addition, these 4 items loaded highly on the factor in the one-factor solution. Therefore, it can be concluded that the test structure was unidimensional and MC and CR items beyond knowledge level measured similar mathematical proficiency.

Table 4-17

Comparison of Factor Loadings of the Three Solutions Based on FIFA[a] (Data Two)

			One-factor[d]	Two-factor[e]		Three- factor		
Item	CL[b]	SC[c]	Factor 1	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3
1. MC2	H	AB	**0.403**[f]	0.175	0.269	**0.398**	0.075	-0.016
2. MC3	C	ME	**0.555**	**0.328**	0.272	**0.522**	0.073	0.012
3. MC4	C	FN	**0.665**	**0.384**	**0.338**	**0.556**	0.074	0.100
4. MC8	H	FN	**0.462**	**0.332**	0.162	**0.413**	-0.004	0.116
5. MC9	H	FN	**0.610**	**0.441**	0.215	**0.547**	0.060	0.064
6. MC13	H	FN	**0.517**	0.199	**0.376**	**0.448**	0.140	0.019
7. MC15	C	FN	**0.358**	0.185	0.203	**0.323**	0.039	0.029
8. MC17	C	FN	**0.475**	0.243	0.280	**0.342**	0.141	0.070
9. CR1	H	FN	**0.496**	0.254	0.296	**0.458**	0.078	0.017
10. CR2	H	ME	**0.694**	**0.550**	**0.179**	**0.796**	-0.097	0.036
11. CR3	H	AB	**0.941**	**1.333**	**-0.464**	-0.026	-0.031	**1.028**
12. CR4	H	AB	**0.962**	**1.314**	**-0.429**	0.016	-0.015	**0.990**
13. CR5	H	FN	**0.532**	**-0.468**	**1.163**	-0.154	**0.982**	0.005
14. CR6	H	FN	**0.518**	**-0.521**	**1.210**	-0.123	**0.989**	-0.051
15. CR7	H	ME	**0.681**	**0.474**	0.254	**0.960**	-0.156	-0.081
16. CR8	H	ME	**0.669**	**0.377**	**0.351**	**0.937**	-0.079	-0.131
Variance			38.25%	39.36%	6.80%	37.78%	8.13%	4.70%
Largest roots			6.03	6.10	1.04	6.18	1.12	0.77
Correlation								
Factor 1				1		1		
Factor 2			-	0.818	1	0.629	1	
Factor 3						0.683	0.393	1

Note. [a]26 items from Booklet 3, Gr.7 and Gr.8, TIMSS (N=2,073). [b]CL-cognitive level, K-nowledge, C – complex procedure, H-higher mental process. [c]SC-sub-content, FN-raction and number sense, ME – measurement, AB – whole number. [d]Factor loadings for one-factor model is principal factor loadings. [e]Factor loadings for two- and three-factor models are promax factor loadings. [f]Bold numbers are loadings greater than the absolute value of 0.30.

Confirmatory Approach

Sixteen items designed to tap high cognitive levels (complex procedures and high mental process) in the TIMSS Mathematics Examination (Gr.7 and Gr.8) were selected for the investigation using MRCMLM. The two-dimension between-item (format) model was compared with the unidimensional model based on the deviance indices of the two models (Table 4-18).

Table 4-18

Comparison of Unidimensional and Two-dimension Format Models Based on MRCMLM (Data Two)

			Unidimensional	Two-dimension	Two-dimension
			MNSQ[d]	Between-item	Within-item
Item[a]	CL[b]	SC[c]	(Wfit[e])	MNSQ (Wfit)	MNSQ (Wfit)
1. MC2	H	AB	1.14 **(4.5)**[f]	1.06 (1.9)	1.05 (1.8)
2. MC3	C	ME	1.02 (0.9)	0.98 (-0.8)	0.97 (-1.2)
3. MC4	C	FN	0.92 **(-2.5)**	0.88 **(-4.0)**	0.89 **(-3.7)**
4. MC8	H	FN	1.12 **(5.2)**	1.04 **(2.1)**	1.07 **(3.4)**
5. MC9	H	FN	1.01 (0.3)	0.98 (-0.8)	0.95 (-1.9)
6. MC13	H	FN	1.07 **(2.5)**	1.04 (1.6)	0.99 (-0.4)
7. MC15	C	FN	**1.20 (6.9)**	1.10 **(3.7)**	1.13 **(5.0)**
8. MC17	C	FN	1.07 **(3.2)**	1.01 (0.5)	1.03 (1.5)
9. CR1	H	FN	1.07 **(2.3)**	1.14 **(4.6)**	**1.21 (5.7)**
10. CR2	H	ME	0.91 **(-3.9)**	0.96 (-1.4)	0.95 (-1.9)
11. CR3	H	AB	0.90 **(-4.5)**	0.88 **(-4.5)**	0.89 **(-4.4)**
12. CR4	H	AB	0.86 **(-5.4)**	0.86 **(-5.2)**	0.86 **(-5.4)**
13. CR5	H	FN	1.01 (0.6)	1.03 (1.5)	1.08 **(3.4)**
14. CR6	H	FN	1.02 (0.9)	1.06 **(2.4)**	1.07 **(2.8)**
15. CR7	H	ME	0.91 **(-4.0)**	0.94 **(-2.4)**	0.96 (-1.4)
16. CR8	H	ME	0.91 **(-3.7)**	0.93 **(-2.4)**	0.96 (-1.4)
Deviance			36934.39	36771.96	36770.71
Estimated parameters			18	21	21
Correlation (dimension)			-	0.910	0.486

Note. [a] 16 items from Booklet 3, Gr.7 and Gr.8, TIMSS (N = 2,073). [b] CL-cognitive level, C-complex procedure, H-higher mental process. [c] SC-sub-content, FN-fraction and number sense, ME – measurement, AB-algebra. [d] MNSQ-mean square residual. [e] Wfit-weighted fit [f] Bold items are misfit.

The two-dimension between-item model was a better model than the unidimensional model, $\chi^2_{(3)} = 162.43$, $p < 0.001$. The large sample size (N = 2,073) may inflate the deviance difference between the models. The correlation between the two dimensions was 0.91, indicating that the two dimensions measured similar ability.

As shown in Table 4-18, there were some misfit items across the three models. The misfit might be due to different discrimination index, guessing, or large sample size. The two-dimension within CR model appeared to be no better than the other two models. The difference of the deviance between the two-dimension between-item model and the two-dimension within-item model was only 1.25, indicating that the CR items did not differ much from the MC items in tapping the mathematical proficiency.

Summary for Research Question Two

The results from the exploratory factor analyses seemed to support the hypothesis that the test structure was unidimensional when the selected items were tapping higher cognitive ability. However, the results from the confirmatory approaches indicated that the two-dimension between-item model was the better model. The examination of factor loadings revealed that the non-significant minor dimensions were due to the item local dependence instead of the format differences. The results from the confirmatory factor analysis should be interpreted with caution because it is possible that the deviance difference was inflated due to the large sample size. In addition, the high correlation (0.91) between the two dimensions (MC and CR items) from MRCMLM seemed to support the findings from FIFA. Therefore, the hypothesis that the test structure was unidimensional was supported.

RESEARCH QUESTION THREE

In order to find out whether high ability students differ from low ability students in dealing with different formats, the following hypotheses were tested:

Hypothesis One: MC and CR items are two dimensions for low ability students.

Hypothesis Two: MC and CR items are one dimension for high ability students.

Hypothesis Three: There is statistically significant interaction between format and ability levels.

Hypothesis One

According to the fit indices in Table 4-19, the two-dimension format model did not fit the data better than the unidimensional model, $\chi^2_{(3)} = 0.97$, $p > 0.05$, indicating that MC and CR items were one dimension for the low ability students. Hypothesis one was rejected.

Table 4-19

Comparison of Unidimensional and Two-dimension Format Models on Low Ability Students' Responses[a] (Data Two)

					Unidimensional		Two-dimension	
Item[b]	CL[c]	SC[d]	Estimate[e]	Error	MNSQ[f]	Wfit[g]	MNSQ	Wfit
1. MC2	H	AB	-0.055	0.081	0.99	-0.5	0.99	-0.7
2. MC3	C	ME	0.703	0.085	0.97	-0.9	1.04	1.3
3. MC4	C	FN	0.362	0.082	0.93	-3.5	0.98	-0.9
4. MC8	H	FN	0.718	0.085	1.00	0.0	0.99	-0.3
5. MC9	H	FN	0.519	0.083	1.04	1.7	0.98	-1.0
6. MC13	H	FN	0.423	0.083	1.01	0.5	1.04	1.9
7. MC15	C	FN	-0.029	0.081	0.98	-1.1	1.03	1.6
8. MC17	C	FN	1.692	0.107	1.02	0.3	1.01	0.2
9. CR1	H	FN	0.208	0.081	1.01	0.5	1.01	0.5
10. CR2	H	ME	3.291	0.201	1.04	0.3	1.00	0.0
11. CR3	H	AB	3.140	0.188	1.04	0.3	0.99	0.0
12. CR4	H	AB	4.182	0.305	1.05	0.3	0.99	0.1
13. CR5	H	FN	1.635	0.105	1.01	0.2	1.00	0.0
14. CR6	H	FN	2.039	0.121	1.01	0.1	0.96	-0.4
15. CR7	H	ME	3.375	0.209	1.04	0.3	1.01	0.1
16. CR8	H	ME	3.213	0.194	1.04	0.3	1.00	0.0
Deviance						9555.23		9554.26
Estimated[h] parameters						18		21
Correlation (dimension)						-		0.545

Note. [a] Low ability-students at the low 30% of the total score (N = 672); [b] 16 items from Booklet 3, Gr.7 and Gr.8, TIMSS. [c] CL-cognitive level, C-omplex procedure, H-higher mental process. [d] SC-sub-content, FN-fraction and number sense, ME-measurement, AB-algebra. [e] Estimate-difficulty parameter, [f] MNSQ-mean square residual. [g]Wfit-weighted fit. [h] the magnitudes are discussed in chapter 5.

Hypothesis Two

According to the deviance indices in Table 4-20, the two-dimension format model seemed to fit the data better than the unidimensional model, $\chi^2_{(3)} = 44.86$, $p < 0.001$, indicating that the test structure was two-dimensional for the high ability group. Hypothesis two was rejected.

Table 4-20

Comparison of Unidimensional and Two-dimension Format Models for High Ability Students' Responses[a] (Data Two)

					Unidimensional		Two-dimension	
Item[b]	CL[c]	SC[d]	Estimate[e]	Error	MNSQ[f]	Wfit[g]	MNSQ	Wfit
1. MC2	H	AB	-2.203	0.124	1.02	0.2	0.99	-0.1
2. MC3	C	ME	-2.173	0.122	0.98	-0.2	0.97	-0.3
3. MC4	C	FN	-3.382	0.201	0.99	-0.0	0.95	-0.2
4. MC8	H	FN	-1.449	0.097	1.02	0.3	0.97	-0.6
5. MC9	H	FN	-2.949	0.166	1.00	0.0	0.96	-0.3
6. MC13	H	FN	-2.383	0.132	1.01	0.1	0.98	-0.2
7. MC15	C	FN	-1.817	0.108	1.03	0.4	0.99	-0.2
8. MC17	C	FN	-0.638	0.083	1.02	0.6	1.00	0.2
9. CR1	H	FN	-2.455	0.136	1.01	0.2	1.04	0.4
10. CR2	H	ME	-0.954	0.087	1.00	0.0	1.06	1.4
11. CR3	H	AB	-0.509	0.082	0.95	-2.0	0.96	-1.4
12. CR4	H	AB	-0.384	0.081	0.93	-3.0	0.94	-2.0
13. CR5	H	FN	-0.814	0.085	1.05	1.5	1.05	1.4
14. CR6	H	FN	-0.814	0.085	1.05	1.7	1.03	1.1
15. CR7	H	ME	-0.509	0.082	0.96	-1.3	1.00	0.0
16. CR8	H	ME	-0.509	0.082	0.99	-0.5	1.00	-0.1
Deviance						10591.69		10546.83
Estimated parameters						18		21
Correlation (dimension)						-		0.408

Note. [a]High ability - students at the low 30% of the total score (N = 676); [b]16 items from Booklet 3, Gr.7 and Gr.8, TIMSS. [c]CL-cognitive level, C-complex procedure, H-higher mental process. [d]SC-sub-content, FN-fraction and number sense, ME-measurement, AB – algebra. [e]stimate-difficulty parameter, [f]MNSQ-mean square residual. [g] Wfit-weighted fit.

Hypothesis Three

Descriptive statistics as well as ANOVA analyses were used to see how high ability students differed from low ability students. According to the item mean scores in Table 4-21, both low and high ability students performed better on MC items than on CR items. However, low ability students seemed to perform much worse on CR items than on MC items in comparison to high ability students.

Table 4-21

Mean Scores of High and Low Ability Students' Responses on MC and CR Items (Data Two)

		Item		
	Ability[a]	Mean	Std. Deviation	N
Multiple-choice	low	0.3475	0.096	672
	high	0.8039	0.084	676
	Total	0.5764	0.2318	1348
Constructed-response	low	0.1142	0.093	672
	high	0.6897	0.1967	676
	Total	0.4028	0.2949	1348

Note. [a]Low ability-students at the low 30% of the total score; High ability-students at the high 30% of the total score.

In Table 4-22, the test statistics showed that the effect of interaction of ability group and format was statistically significant, $F_{(1, 1346)} = 139.43$, $p < 0.001$. The effect size f of the interaction was 0.322, which was medium according to Cohen (1988, p.286). Therefore, it is possible to conclude that high ability students differed from low ability students to some extent in dealing with MC and CR items (see also Figure 10). Research hypothesis three was supported.

Table 4-22

Analysis of Variance for High and Low Ability Students' Mean Scores on MC and CR Items (Data Two)

Source	SS	df	MS	F	Eta (η^2)	f
Between			Between			
Ability Group	179.437	1	179.437	9377.174*	0.874	2.634
Error (Group)	25.756	1346	0.019			
Within			Within			
Format[a]	20.342	1	20.342	1187.308*	0.469	1.381
Format * Group[b]	2.389	1	2.389	139.428*	0.094	0.322
Error (Format)	20.060	1346	0.017			

Note. [a] Format-MC and CR; [b] Low ability-students at the low 30% of the total score (N = 672); High ability-students at the high 30% of the total score (N = 676). * $p < .001$.

Summary of Research Question Three

The results from the confirmatory approaches indicated that, for low ability students, the unidimensional model was the better model, indicating that the two formats were not distinct constructs. Hypothesis one was rejected. However, for high ability students, the two-dimension format model was better than the unidimensional model. Hypothesis two was rejected.

In terms of hypothesis three, the fact that the test structure was two-dimensional for high ability students and unidimensional for low ability students indicated that different groups of students dealt with MC and CR items differently. In addition, based on the ANOVA analysis, there was statistically significant interaction between ability and format, indicating that the two groups did differ in dealing with different formats (MC vs. CR). Hypothesis three was supported.

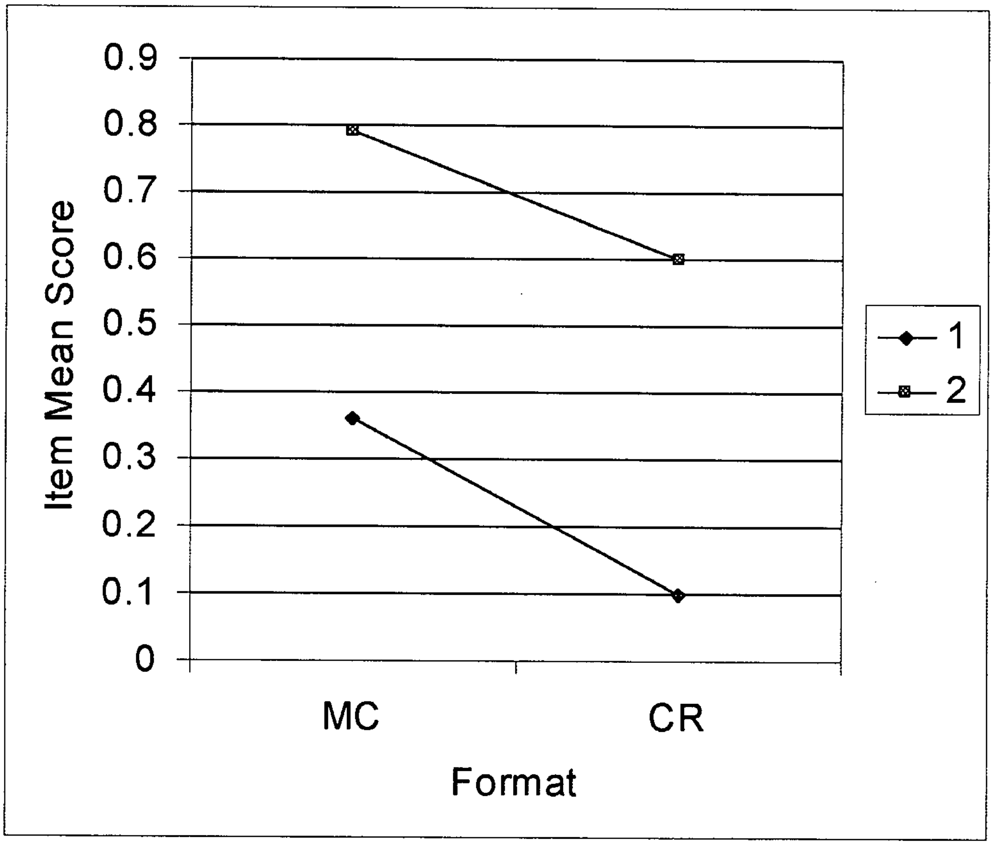


Figure 10: Comparison of High and Low Ability Students in terms of Format Differences (Data Set Two). MC – Multiple-Choice, CR – Constructed-Response, 1 – Low ability, 2 – High ability

EXAMINATION OF DATA SET THREE:
BRITISH COLUMBIA EXAMINATION (APRIL 1998)

Missing Data

The analyses were based on the British Columbia Grade 12 Mathematics Examination (April 1998), from which 23 items were selected for the investigation. The sample (N = 1,718) available from the British Columbia Grade 12 Mathematics Examination data bank was used. As shown in Table 4-23, only CR items were reported to have missing responses in the data bank. The average response rates for CR items were 94.8%. Speededness was not a problem because no item had more than 10% of the missing response at the end of the test.

Table 4-23

Percentage of Missing Data (Data Set Three)

Item	Valid Response Rate	Percent of Missing Response
1. MC1	100.0	0.0
2. MC2	100.0	0.0
3. MC3	100.0	0.0
4. MC4	100.0	0.0
5. MC5	100.0	0.0
6. MC6	100.0	0.0
7. MC7	100.0	0.0
8. MC8	100.0	0.0
9. MC9	100.0	0.0
10. MC10	100.0	0.0
11. MC11	100.0	0.0
12. MC12	100.0	0.0
13. MC13	100.0	0.0
14. MC14	100.0	0.0
15. MC15	100.0	0.0
16. MC16	100.0	0.0
17. MC17	100.0	0.0
18. MC18	100.0	0.0
Mean	100.0	0.0
19. CR1	93.9	6.1
20. CR2	98.8	1.2
21. CR3	93.4	6.6
22. CR4	96.8	3.2
23. CR5	**91.2**	**8.8**
Mean	94.8	5.2

Note. 23 items from BC Grade 12 Mathematics Examination (April 1998). Sample size is 1,718.

Constructed-Response Items

There were 5 CR items that were coded at multi-levels (maximum 4 levels). . For the four-point open-ended response items, scores of zero, one and two were re-coded to a value of zero, while score of two and three were re-coded to a value of one. The information function was plotted to see if there was any information loss after dichotomizing the CR items. The information loss was 13% (see Figure 11, Appendix C). It seemed that the loss of information might affect the results from FIFA and MRCMLM to some extent. Therefore, the conclusions should be interpreted with caution.

Results from Classical Test Theory

From Table 4-24, it can be seen that students scored higher on MC items than on CR items (item mean score for MC items was 0.74; item mean score for CR items was 0.63). Based on the item total correlation, it seemed that they varied a lot from 0.16 to 0.48. On average, CR items had higher item total correlation than MC items. The overall reliability (Cronbach's alpha) was 0.78.

Table 4-24

Descriptive Item Analysis[a] (Data Set Three)

Item	CL[b]	SC[c]	Item Mean	Std Dev	Item Total Correlation[d]
1. MC1	K	QR	0.93	0.25	0.16
2. MC2	U	QR	0.79	0.41	0.36
3. MC3	H	QR	0.53	0.50	0.39
4. MC4	K	TR	0.93	0.26	0.20
5. MC5	U	TR	0.83	0.38	0.31
6. MC6	H	TR	0.51	0.50	0.37
7. MC7	K	EL	0.94	0.23	0.30
8. MC8	U	EL	0.96	0.21	0.22
9. MC9	H	EL	0.50	0.50	0.47
10. MC10	K	PF	0.87	0.34	0.28
11. MC11	U	PF	0.83	0.37	0.27
12. MC12	H	PF	0.74	0.44	0.35
13. MC13	U	SS	0.92	0.26	0.18
14. MC14	K	SS	0.70	0.46	0.30
15. MC15	H	SS	0.30	0.46	0.20
16. MC16	K	IC	0.93	0.26	0.23
17. MC17	U	IC	0.68	0.47	0.48
18. MC18	H	IC	0.43	0.50	0.35
19. CR1	U	TR	0.57	0.50	0.43
20. CR2	U	EL	0.85	0.36	0.43
21. CR3	U	PF	0.45	0.50	0.34
22. CR4	U	QR	0.84	0.36	0.28
23. CR5	U	EL	0.44	0.50	0.38
Mean for MC			0.74		
Reliability [e]			0.72		
Mean for CR			0.63		
Reliability			0.57		
Mean			0.72		
Reliability for MC and CR			0.78		

Note. [a] 23 items from BC Grade 12 Mathematics Exam (April 1998, N = 1,718). [b] CL-cognitive level, K-knowledge, U-understanding, H-higher mental process. [c] SC-sub-content, TR-trigonometry, QR-Quadratic Relations, EL-Exponential Logarithmic Functions, PF-Polynomial Functions, SS-Sequences and Series, IC-Introduction to Calculus. [d] Discrimination index: point-biserial. [e] Reliability – Cronbach's alpha.

RESEARCH QUESTION ONE

Exploratory Factor Analysis

Three models were tested to investigate whether the format differences affect the dimensionality of the test structure.

Hypothesis one: The test structure is unidimensional.

Hypothesis two: The test structure is two-dimensional.

Hypothesis three: The test structure is three-dimensional.

FIFA was applied in an exploratory approach to test the hypotheses. By looking at the loadings in the one-factor model in Table 4-25, it seemed that all items loaded on the factor. The total variance accounted for by the model was 29.3%, and the largest root was 6.23. Hypothesis one was supported. In terms of the two-factor model, the first dimension was significant ($\lambda = 6.27$), but the second dimension was not ($\lambda = 1.02$). Thus, hypothesis two was rejected. Similarly, the first dimension of the three-factor model was significant ($\lambda = 6.30$), but the second ($\lambda = 1.05$) and third factors ($\lambda = 0.57$) were not. Therefore, hypothesis three was rejected.

Examination of Factor Loadings

Based on the factor loadings of the one-factor solution (see Table 4-25), all MC and CR items loaded on the mathematical proficiency factor with loadings greater than 0.30, indicating that the test structure was unidimensional.

The minor dimensions in the two-factor solution seemed to be represented by three CR items (#1, #3 and #5). In terms of the MC items, no MC items loaded highly on the minor dimension. The 3 CR items representing the minor dimension requested higher cognitive demand than the other items because more procedures seemed to be involved. For CR #3 and #5, first of all, students needed to understand conceptually the problem using their existing knowledge; second, they needed to use schema knowledge to correctly represent the problem structure; third, reasoning and computation skills were needed to produce the correct answer. CR #1 seemed to be less complicated than the other two items. Students needed to apply step one and three only. The mean scores of these 3 items were lower than most of the other items (0.57, 0.45, and 0.44), indicating that students' performances on these two items were poor compared to their performances on the other items. Because the minor dimension represented by the 3 CR items was trivial based on the variance explained and root criteria, the test structure can still be considered to be unidimensional. The MC and CR items seemed to measure similar mathematical proficiency.

Table 4-25

Comparison of Factor Loadings of the Three Solutions Based on FIFA[a] (Data Three)

			One-factor[d]	Two-factor[e]		Three-factor		
Item	CL[b]	SC[c]	Factor 1	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3
1.MC1	K	QR	**0.381**[f]	**0.504**	-0.122	**0.429**	-0.037	0.081
2.MC2	U	QR	**0.568**	**0.585**	0.011	**0.488**	0.086	0.141
3.MC3	H	QR	**0.558**	**0.415**	0.183	0.299	0.195	0.188
4.MC4	K	TR	**0.440**	**0.529**	-0.084	**0.516**	0.044	0.000
5.MC5	U	TR	**0.515**	**0.431**	0.120	0.165	0.054	**0.375**
6.MC6	H	TR	**0.533**	0.237	**0.342**	-0.090	0.161	**0.502**
7.MC7	K	EL	**0.718**	**0.766**	-0.035	0.281	-0.116	**0.630**
8.MC8	U	EL	**0.557**	**0.747**	-0.207	0.228	-0.298	**0.646**
9.MC9	H	EL	**0.665**	**0.344**	**0.377**	-0.007	0.202	**0.540**
10.MC10	K	PF	**0.513**	**0.634**	-0.117	0.215	-0.167	**0.510**
11.MC11	U	PF	**0.451**	**0.355**	0.123	0.089	0.046	**0.369**
12.MC12	H	PF	**0.544**	**0.414**	0.163	0.094	0.053	**0.457**
13.MC13	U	SS	**0.401**	0.202	0.235	0.138	0.212	0.131
14.MC14	K	SS	**0.457**	0.273	0.218	0.081	0.142	0.296
15.MC15	H	SS	**0.310**	0.210	0.123	-0.034	0.020	**0.340**
16.MC16	K	IC	**0.489**	**0.628**	-0.131	**0.556**	-0.018	0.075
17.MC17	U	IC	**0.705**	**0.530**	0.224	0.211	0.132	**0.467**
18.MC18	H	IC	**0.506**	**0.417**	0.123	-0.042	-0.063	**0.634**
19.CR1	U	TR	**0.609**	0.073	**0.615**	**0.351**	**0.779**	-0.282
20.CR2	U	EL	**0.745**	**0.605**	0.184	**0.438**	0.202	0.259
21.CR3	U	PF	**0.500**	-0.245	**0.836**	-0.084	**0.694**	0.005
22.CR4	U	QR	**0.494**	0.258	0.279	0.150	0.247	0.192
23.CR5	U	EL	**0.562**	-0.108	**0.754**	-0.110	**0.578**	0.188
Variance			29.32%	29.09%	3.62%	29.05%	3.81%	2.38%
largest roots			6.23	6.27	1.02	6.30	1.05	0.57
Correlation								
Factor1				1		1		
Factor2			-	0.713	1	0.395	1	
Factor3						0.574	0.679	1

Note. [a] 23 items from BC Grade 12 Mathematics Exam (April 1998, N = 1,718). [b] CL-cognitive level, K-knowledge, U-understanding, H-higher mental process. [c] SC-sub-content, TR-trigonometry, QR-Quadratic Relations, EL-Exponential Logarithmic Functions, PF-Polynomial Functions, SS-Sequences and Series, IC-Introduction to Calculus. [d] Factor loadings for one-factor model are principal factor loadings. [e] Factor loadings for two- and three-factor models are promax factor loadings. [f] Bold numbers are loadings greater than the absolute value of 0.30.

CR item#1

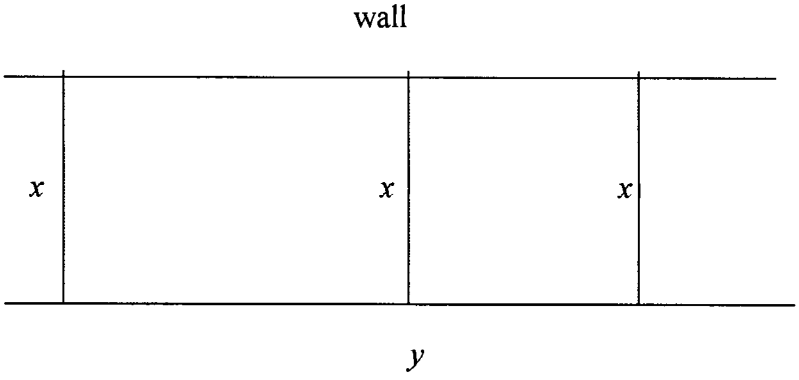
Solve for x: 3 sin2 x – 8sin x + 4 = 0, where $0 \le 2\pi$. (Accurate to at least 2 decimal places.)

CR item#3

Determine the polynomia function of degree 3, with zeros of –2, 0, and 3 that passes through the point (2, 5). Answer may be left in factored form.

CR item#5

A rectangular pigen is to be constructed having one side along an existing wall. The pigpen is also to be divided into two parts as shown in the diagram.



If a total of 300 metres of fencing is used, determine the maximum area that the pigpen can have.

The G^2 index (see Table 4-26) was not used to judge the significance of the dimensionality because it was inflated by the large sample size (N = 1718). The two- and three-factor models were no better models than the one-factor model based on the loadings, variance and root criteria. Therefore, the multiple evidence suggested that the one-factor model was the best model to be accepted.

Table 4-26

Change of the Likelihood Ratio G^2 in the Item Factor Analysis

Factor	G^2	df	p
2 vs. 1	146.34	22	0.000
3 vs. 2	64.62	21	0.000

Confirmatory Approach

The two-dimension format model was compared with the unidimensional model using MRCMLM. Because the two models were hierarchical, the Chi-square difference test was used to compare the models based on the deviance index. According to the deviance indices in Table 4-27, it seemed that the unidimensional model fit the data better than the two-dimension format model, $\chi^2_{(3)} = 2259.70$, $p < 0.001$. The correlation between the two dimensions was 0.82, indicating that the two dimensions measured similar ability.

Based on the weighted fit indices, it seemed that both models did not fit the data very well. For the unidimensional model, there were 5 misfit items. However, no item was misfit item according to MNSQ. For the two-dimension format model, 10 out of 23 items were misfit items, and 5 of them were serious misfit items, however, 4 items were misfit items based on MNSQ. Therefore, it is possible to conclude that the unidimensional model was much better than the two-dimension format model

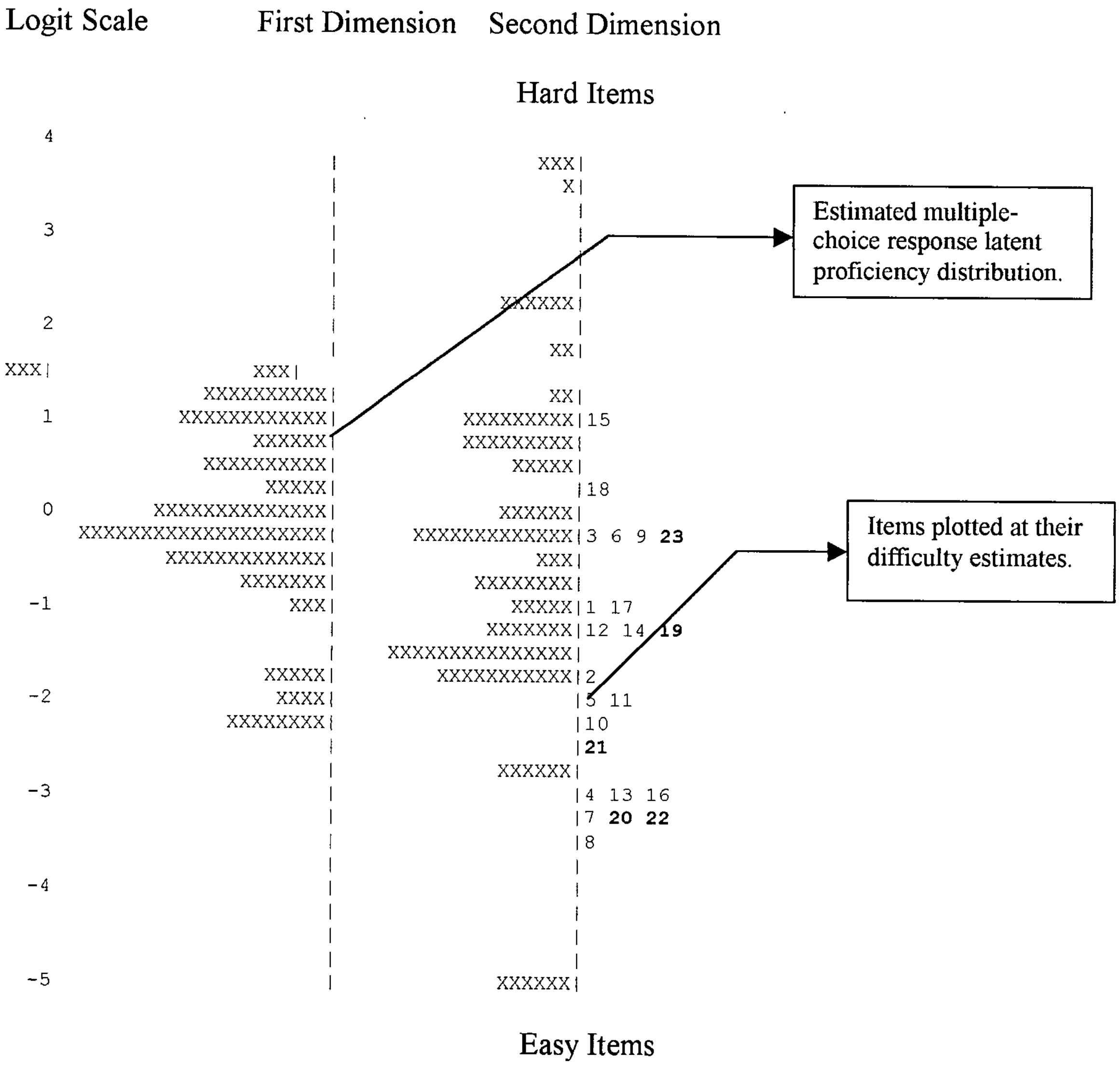
Table 4-27

Comparison of Unidimensional and Two-dimension Format Models Based on MRCMLM[a] (Data Three)

					Unidimensional		Two-dimension	
Item	CL[b]	SC[c]	Estimate[d]	Error	MNSQ[e]	Wfit[f]	MNSQ	Wfit
1.MC1	K	QR	-3.098	0.100	1.03	0.4	**0.80**	**-7.9**[g]
2.MC2	U	QR	-1.621	0.064	0.98	-0.4	0.96	-1.2
3.MC3	H	QR	-0.162	0.054	0.99	-0.3	1.00	-0.2
4.MC4	K	TR	-2.992	0.096	1.01	0.2	0.98	-0.3
5.MC5	U	TR	-1.897	0.068	1.02	0.5	1.00	0.1
6.MC6	H	TR	-0.055	0.054	0.99	-0.3	1.00	-0.0
7.MC7	K	EL	-3.259	0.107	0.90	-1.2	0.86	-1.8
8.MC8	U	EL	-3.566	0.121	0.95	-0.5	0.90	-1.1
9.MC9	H	EL	0.001	0.054	0.94	**-3.3**	0.92	**-4.1**
10.MC10	K	PF	-2.280	0.076	1.03	0.6	1.01	0.3
11.MC11	U	PF	-1.944	0.069	1.06	1.4	1.03	0.8
12.MC12	H	PF	-1.266	0.060	1.02	0.7	1.03	0.9
13.MC13	U	SS	-2.955	0.095	1.02	0.3	1.02	0.2
14.MC14	K	SS	-1.045	0.058	1.08	**2.8**	1.07	**2.3**
15.MC15	H	SS	1.054	0.059	1.09	**3.6**	1.08	**3.1**
16.MC16	K	IC	-3.001	0.097	0.97	-0.4	0.95	-0.7
17.MC17	U	IC	-0.959	0.057	0.90	**-3.7**	0.93	**-2.7**
18.MC18	H	IC	0.344	0.055	1.01	0.6	1.01	0.5
19.CR1	U	TR	-0.370	0.054	0.97	-1.4	1.12	**4.3**
20.CR2	U	EL	-2.104	0.072	0.90	**-2.2**	**1.19**	**3.0**
21.CR3	U	PF	0.217	0.054	1.03	1.5	**3.16**	**31.9**
22.CR4	U	QR	-2.048	0.071	1.07	1.5	**1.46**	**6.8**
23.CR5	U	EL	0.302	0.054	0.99	-0.6	1.11	**3.9**
Deviance						35927.44		38187.14
Estimated[h] parameters						25		28
Correlation (dimension)						-		0.819

Note. [a] 23 items from BC Grade 12 Mathematics Examination (April 1998, N =1,780). [b] CL-cognitive level, K-knowledge, U understanding, H – higher mental process. [c]SC-sub-content, TR-trigonometry, QR-Quadratic Relations, EL-Exponential Logarithmic Functions, PF-Polynomial Functions, SS-Sequences and Series, IC-Introduction to Calculus. [d] Estimate – difficulty parameter, [e] MNSQ – mean square residual. [f] Wfit-weighted fit. [g] Bold indicates misfit items.
[h] the magnitudes are discussed in chapter 5.

Based on the latent distribution of the two-dimension format model (Figure 13), the first dimension (MC items) looked normally distributed whereas the second dimension (CR items) was slightly skewed negatively. The students' distribution of the second dimension spread around more than the distribution of the first dimension.



```
Logit Scale          First Dimension    Second Dimension

                                          Hard Items
  4
                              |                     XXX|
                              |                       X|
                              |                        |
  3                           |                        |
                              |                        |
                              |                        |
                              |                  XXXXXX|
  2                           |                        |
                              |                      XX|
XXX|                   XXX|
                 XXXXXXXXXXX|                        XX|
  1            XXXXXXXXXXXXX|               XXXXXXXXX|15
                      XXXXXX|               XXXXXXXXX|
                 XXXXXXXXXXX|                   XXXXX|
                       XXXXX|                        |18
  0          XXXXXXXXXXXXXXX|                  XXXXXX|
      XXXXXXXXXXXXXXXXXXXXXX|           XXXXXXXXXXXXX|3  6  9  23
              XXXXXXXXXXXXXX|                     XXX|
                     XXXXXXX|                XXXXXXXX|
 -1                      XXX|                   XXXXX|1  17
                            |                 XXXXXXX|12  14  19
                            |         XXXXXXXXXXXXXXX|
                       XXXXX|             XXXXXXXXXXX|2
 -2                     XXXX|                        |5  11
                    XXXXXXXX|                        |10
                            |                        |21
                            |                  XXXXXX|
 -3                         |                        |4  13  16
                            |                        |7  20  22
                            |                        |8
                            |                        |
 -4                         |                        |
                            |                        |
                            |                        |
                            |                        |
 -5                         |                  XXXXXX|

                                          Easy Items
```

Estimated multiple-choice response latent proficiency distribution.
Items plotted at their difficulty estimates.

Figure 13. Map of Latent Distribution and Response Model Parameter Estimates for Two-dimension Format Model. Bold items are constructed-response items, non-bold items are multiple-choice items (23 items from BC Grade 12 Mathematics Examination, April 1998).

Summary for Research Question One

The results from FIFA and MRCMLM indicated that the data structure was unidimensional. The evidence from FIFA revealed that the test structure was dominated by one factor. All MC and CR items loaded on the first factor in the one-factor solution with loadings greater than 0.30. The examination of the 3 CR items loading on the minor dimension in the two-factor solution indicated that they had higher cognitive demand upon students. However, the minor dimension represented by the 3 CR items was not significant based on the variance explained and root criteria. The results from MRCMLM also indicated that the unidimensional model was much better than the two-dimension format model. The hypothesis that MC and CR items represented two dimensions was not supported.

RESEARCH QUESTION TWO

Exploratory Factor Analysis

Three hypotheses were tested to investigate whether MC and CR items differ in measuring students' cognitive ability beyond knowledge level in mathematics.

Hypothesis One: The test structure is unidimensional.

Hypothesis Two: The test structure is two-dimensional.

Hypothesis Three: The test structure is three-dimensional.

FIFA was applied as an exploratory approach to test the hypotheses. Seventeen items designed to tap high cognitive levels (understanding and high mental process) in the British Columbia Grade12 Mathematics Examination (April) were selected for the investigation. According to the variance, and root criteria in Table 4-28, it seemed that the one-factor model was the best of all the three models. The total variance accounted for the model was 30.2%, and the largest root was 4.86, which was greater than 1.4, indicating that the first factor was a significant factor. In terms of the two-factor model, the first factor was significant ($\lambda = 4.91$), but the second factor was not ($\lambda = 0.75$). Thus, hypothesis two was not supported. Similarly, the first factor was significant in the three-factor solution ($\lambda = 4.94$), but the second ($\lambda = 0.76$) and third factors ($\lambda = 0.41$) were not. Therefore, hypothesis three was rejected.

Examination of Factor Loadings

By looking at the one-factor solution, all the MC and CR items loaded on the factor with loadings greater then 0.30. Those MC and CR items beyond knowledge level seemed to measure same mathematical proficiency. The minor dimension in the two-factor solution was represented by 3 CR items. The examination of the above items showed that they requested higher cognitive ability upon students and students' performances on them were not satisfactory. The fact that the other CR items did not seem to measure higher cognitive ability than the MC items indicated that the cognitive demand instead of the format differences was the reason for the minor dimensions. In addition, the minor dimension was trivial in comparison to the first dimension in terms of the variance and root criteria. Therefore, MC and CR items can be considered to measure similar mathematical proficiency.

Table 4-28

Comparison of Factor Loadings of the Three Solutions Based on FIFA[a] (Data Three)

			One-factor[d]	Two- factor[e]		Three-factor		
Item	CL[b]	SC[c]	Factor 1	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3
1.MC2	U	QR	**0.548**	**0.561**	0.015	0.099	-0.002	**0.564**[f]
2.MC3	H	QR	**0.548**	**0.417**	0.165	0.185	0.138	**0.331**
3.MC5	U	TR	**0.515**	**0.521**	0.022	**0.363**	-0.001	0.221
4.MC6	H	TR	**0.531**	0.276	0.293	**0.473**	0.208	-0.089
5.MC8	U	EL	**0.517**	**0.871**	**-0.360**	**0.563**	**-0.333**	**0.302**
6.MC9	H	EL	**0.672**	**0.430**	0.287	**0.558**	0.196	0.001
7.MC11	U	PF	**0.446**	**0.397**	0.071	0.292	0.042	0.170
8.MC12	H	PF	**0.551**	**0.570**	0.008	**0.502**	-0.022	0.129
9.MC13	U	SS	**0.398**	0.243	0.183	0.251	0.150	0.059
10.MC15	H	SS	**0.312**	0.292	0.039	**0.322**	0.017	0.001
11.MC17	U	IC	**0.688**	**0.530**	0.200	**0.443**	0.141	0.201
12.MC18	H	IC	**0.497**	**0.498**	0.027	**0.601**	-0.028	-0.039
13.CR1	U	TR	**0.624**	0.036	**0.660**	-0.231	**0.653**	**0.422**
14.CR2	U	EL	**0.723**	**0.638**	0.124	0.211	0.095	**0.545**
15.CR3	U	PF	**0.533**	-0.187	**0.799**	0.047	**0.625**	0.001
16.CR4	U	QR	**0.496**	0.291	0.239	0.076	0.203	**0.317**
17.CR5	U	EL	**0.585**	-0.060	**0.716**	**0.254**	**0.578**	-0.127
Variance			30.18%	30.42%	3.59%	30.29%	3.86%	2.18%
largest roots			4.86	4.91	0.75	4.94	0.76	0.41
Correlation								
Factor 1				1		1		
Factor 2				0.747	1	0.613	1	
Factor 3						0.657	0.460	1

Note. [a] 17 items from BC Grade12 Mathematics Examination (April 1998, N =1,780). [b] CL-cognitive level, U-understanding, H-higher mental process. [c] SC-sub-content, TR-trigonometry, QR-Quadratic Relations, EL-Exponential Logarithmic Functions, PF-Polynomial Functions, SS-Sequences and Series, IC-Introduction to Calculus. [d] Factor loadings for one-factor model are principal factor loadings. [e] Factor loadings for two- and three-factor models are promax factor loadings. [f] Bold numbers are loadings greater than the absolute value of 0.30.

Confirmatory Approach

The two-dimension format model was compared with the unidimensional model using MRCMLM (see Table 4-29). The two-dimension model was significantly better than the unidimensional model, $\chi^2_{(3)} = 46.13$, $p < 0.001$. The significant deviance difference may be due to large sample size. The correlation between the two dimensions was 0.86, indicating that the two dimensions measured similar mathematical ability.

Because there was no deviance difference between the two-dimension within-item model and two-dimension between-item model, we can conclude that CR items did not tap additional construct that was different from the overall mathematical ability.

Table 4-29

Comparison of Unidimensional and Two-dimension Format Models Based on MRCMLM[a] (Data Three)

Item	CL[b]	SC[c]	Unidimensional MNSQ[d] (Wfit[e])	Two-dimension Between-item MNSQ (Wfit)	Two-dimension Within-item MNSQ (Wfit)
1.MC2	U	QR	1.01(0.2)	1.02 (0.5)	0.98 (-0.5)
2.MC3	H	QR	1.00 (0.1)	0.99 (-0.5)	1.01 (0.4)
3.MC5	U	TR	1.02 (0.4)	1.04 (0.9)	1.01 (0.4)
4.MC6	H	TR	1.02 (1.0)	1.03 (1.6)	1.02 (0.9)
5.MC8	U	EL	0.95 (-0.5)	0.97 (-0.2)	0.93 (-0.8)
6.MC9	H	EL	0.93 **(-3.4)**	0.93 **(-3.5)**	0.96 **(-2.2)**
7.MC11	U	PF	1.05 (1.1)	1.09 (1.9)	1.06 (1.4)
8.MC12	H	PF	1.05 (1.7)	1.05 (1.7)	1.05 (1.4)
9.MC13	U	SS	1.02 (0.3)	1.04 (0.5)	1.01 (0.1)
10.MC15	H	SS	1.09 **(3.7)**	1.09 **(3.5)**	1.11 **(4.2)**
11.MC17	U	IC	0.94 **(-2.2)**	0.93 **(-2.7)**	0.94 **(-2.1)**
12.MC18	H	IC	1.04 (1.9)	1.02 (0.8)	1.03 (1.7)
13.CR1	U	TR	0.95 **(-2.3)**	0.99 (-0.5)	0.99 (-0.6)
14.CR2	U	EL	0.93 (-1.5)	1.05 (1.0)	0.97 (-0.6)
15.CR3	U	PF	1.05 **(2.3)**	1.02 (0.9)	1.02 (0.9)
16.CR4	U	QR	1.05 (1.1)	**1.19 (3.6)**	1.11 **(2.4)**
17.CR5	U	EL	0.97 (-1.6)	1.01 (0.6)	1.03 (1.1)
Deviance			29690.59	29644.46	29644.35
Estimated parameters			19	22	22
Correlation (correlation)			-	0.860	0.149

Note. [a] 17 items from BC Grade12 Mathematics Examination (April 1998, N = 1,780). [b] CL-cognitive level, U-understanding, H-higher mental process. [c] SC-sub-content, TR-trigonometry, QR-Quadratic Relations, EL-Exponential Logarithmic Functions, PF-Polynomial Functions, SS-Sequences and Series, IC-Introduction to Calculus. [d] MNSQ-mean square residual. [e] Wfit-weighted fit.

Summary for Research Question Two

The results from the exploratory factor analyses seemed to support the hypothesis that the test structure was unidimensional when selected items were tapping similar cognitive ability beyond knowledge level. The examination of the 3 CR items that loaded on the non-significant minor dimension in the two-factor solution indicated that they might tap higher cognitive ability. However, the other two CR items did not seem to tap higher cognitive ability than the MC items. Therefor, format differences did not seem to be the reason for the minor dimension. The results from the confirmatory approaches indicated that the two-dimension between-item format model was the better model. The results from the confirmatory factor analysis should be interpreted with caution because it is possible that the deviance difference was inflated due to the large sample size. The medium high correlation between the two dimensions (MRCMLM) confirmed that the two dimensions measured similar constructs. In addition, the two-dimension within-item model was no better than the two-dimension between-item model, indicating that CR and MC items measured similar ability. Therefore, the hypothesis that the test structure was unidimensional was supported.

RESEARCH QUESTION THREE

To test the hypothesis that whether high ability students differ from low ability students in dealing with different formats, the following hypotheses were tested:

Hypothesis One: MC and CR items are two dimensions for low ability students.

Hypothesis Two: MC and CR items are one dimension for high ability students.

Hypothesis Three: There is statistically significant interaction between format and ability levels.

Hypothesis One

According to the deviance indices in Table 4-30, the two-dimension format model seemed to fit the data better than the unidimensional model for the low ability students, $\chi^2_{(3)} = 28.97$, $p < 0.001$. Hypothesis one was supported.

Table 4-30

Comparison of Unidimensional and Two-dimension Format Models on Low Ability Students' Responses[a] (Data Three)

					Unidimensional		Two-dimension	
Item[b]	CL[c]	SC[d]	Estimate[e]	Error	MNSQ[f]	Wfit[g]	MNSQ	Wfit
1.MC2	U	QR	-0.610	0.073	0.99	-0.5	0.98	-0.8
2.MC3	H	QR	0.839	0.076	1.01	0.5	1.03	0.9
3.MC5	U	TR	-0.925	0.077	0.97	-0.8	0.96	-1.3
4.MC6	H	TR	0.909	0.077	1.03	0.8	1.02	0.7
5.MC8	U	EL	-2.472	0.126	0.99	-0.1	0.99	-0.1
6.MC9	H	EL	1.171	0.081	1.00	0.0	1.01	0.2
7.MC11	U	PF	-0.992	0.078	1.00	0.1	0.99	-0.4
8.MC12	H	PF	-0.224	0.071	1.01	0.7	1.02	1.4
9.MC13	U	SS	-2.009	0.106	0.99	-0.1	0.99	-0.2
10.MC15	H	SS	1.511	0.089	1.04	0.7	1.01	0.3
11.MC17	U	IC	0.212	0.071	1.00	-0.1	0.99	-0.7
12.MC18	H	IC	1.198	0.082	1.03	0.8	1.01	0.2
13.CR1	U	TR	0.766	0.075	1.00	-0.1	0.97	-1.0
14.CR2	U	EL	-0.949	0.078	0.97	-1.0	0.97	-0.7
15.CR3	U	PF	1.145	0.081	1.02	0.5	1.00	0.1
16.CR4	U	QR	-1.042	0.079	1.00	0.1	1.00	0.1
17.CR5	U	EL	1.245	0.083	1.00	-0.0	0.99	-0.2
Deviance						16097.46		16068.49
Estimated parameters						19		22
Correlation (correlation)						-		0.129

Note. [a] Low ability-students at the low 40% of the total score (N = 840); [b] 17 items from BC Grade12 Mathematics Examination (April 1998). [c] CL-cognitive level, U-understanding, H-higher mental process, [d]C-sub-content, TR-trigonometry, QR-Quadratic Relations, EL-Exponential Logarithmic Functions, PF-Polynomial Functions, SS-Sequences and Series, IC-Introduction to Calculus, [e]Estimate – difficulty parameter, [f]MNSQ – mean square residual, [g]Wfit-weighted fit.

Hypothesis Two

According to the deviance indices in Table 4-31, the two-dimension format model and the unidimensional model seemed to fit the data similarly for the high ability students, $\chi^2_{(3)} = 11.77$, $p > 0.05$, indicating that MC and CR items measured similar construct for the high ability students. Hypothesis two was supported.

Table 4-31

Comparison of Unidimensional and Two-dimension Format Models on High Ability Students' Responses[a] (Data Three)

					Unidimensional		Two-dimension	
Item[b]	CL[c]	SC[d]	Estimate[e]	Error	MNSQ[f]	Wfit[g]	MNSQ	Wfit
1.MC2	U	QR	-2.789	0.155	0.99	-0.1	1.00	0.0
2.MC3	H	QR	-1.259	0.090	0.97	-0.6	1.00	0.0
3.MC5	U	TR	-2.915	0.163	0.99	-0.0	0.98	-0.1
4.MC6	H	TR	-1.195	0.088	1.01	0.3	1.00	-0.1
5.MC8	U	EL	-5.567	0.579	0.96	0.1	0.98	0.2
6.MC9	H	EL	-1.375	0.093	0.98	-0.3	1.01	0.2
7.MC11	U	PF	-2.915	0.163	0.96	-0.2	1.01	0.1
8.MC12	H	PF	-2.653	0.146	0.96	-0.3	0.99	-0.1
9.MC13	U	SS	-4.085	0.280	0.96	-0.1	0.98	0.0
10.MC15	H	SS	0.241	0.076	1.03	1.9	0.99	-0.3
11.MC17	U	IC	-2.612	0.144	0.95	-0.4	0.99	-0.0
12.MC18	H	IC	-0.783	0.081	1.00	-0.2	1.03	0.9
13.CR1	U	TR	-1.736	0.103	0.98	0.3	1.03	0.5
14.CR2	U	EL	-4.578	0.356	0.96	0.0	0.88	-0.3
15.CR3	U	PF	-0.945	0.084	1.00	-0.1	1.08	1.8
16.CR4	U	QR	-3.255	0.190	0.97	-0.1	0.94	-0.3
17.CR5	U	EL	-0.877	0.082	0.98	-0.5	0.97	-0.8
Deviance						8794.67		8782.90
Estimated parameters						19		22
Correlation (correlation)						-		0.242

Note. [a]High ability - students at the high 40% of the total score (N = 724); [b]17 items from BC Grade12 Mathematics Examination (April 1998). [c]CL-cognitive level, U- understanding, H-higher mental process. [d] SC-sub-content, TR-trigonometry, QR-Quadratic Relations, EL-Exponential Logarithmic Functions, PF-Polynomial Functions, SS-Sequences and Series, IC-Introduction to Calculus. [e] Estimate-difficulty parameter, [f] MNSQ-mean square residual. [g]Wfit-weighted fit.

Hypothesis Three

Descriptive statistics as well as ANOVA analyses were used to see how high ability students differed from low ability students. According to the item mean scores in Table 4-32, the low ability group performed worse on CR items than on MC items. However, it was not the case for the high ability group. High ability students performed similarly on both formats.

Table 4-32

Mean Scores of High and Low Ability Students' Responses on MC and CR Items (Data Three)

		Item		
	Ability[a]	Mean	Std. Deviation	N
Multiple-choice	low	0.5126	0.1399	840
	high	0.8420	0.1023	724
	Total	0.6651	0.2057	1564
Constructed-response	low	0.4505	0.2221	840
	high	0.8406	0.1701	724
	Total	0.6311	0.2788	1564

Note. [a] Low ability - students at the low 40% of the total score (N = 840); High ability students at the high 40% of the total score (N=724).

In Table 4-33, the test statistics showed that the interaction between ability group and format was statistically significant, $F_{(1, 1562)} = 27.01$, $p < 0.001$, see also Figure14. The effect size for the interaction was only 0.132, which was small (Cohen, 1988, p.285). Therefore, it is possible that the statistical significance of the interaction may be due to other factors such as large sample size. The conclusion should be interpreted with caution.

Table 4-33

Analysis of Variance for High and Low Ability Students' Mean Scores on MC and CR Items (Data Three)

Source	SS	df	MS	F	Eta (η^2)	f
Between			Between			
Group	100.649	1	100.649	3511.493*	0.692	2.247
Error (Group)	44.771	1562	0.029			
Within			Within			
Format[a]	0.784	1	0.784	29.480*	0.019	0.139
Format * Group[b]	0.718	1	0.718	27.011*	0.017	0.132
Error (Format)	41.514	1562	0.027			

Note. [a] Format-MC and CR; [b] Low ability-students at the low 40% of the total score (N=840); High ability-students at the high 40% of the total score (N=724). * $p < .001$.

Summary of Research Question Three

The results from the confirmatory approaches revealed that, for low ability students, the two-dimension between-item format model was a better model, indicating that the two formats might be somewhat different. However, for high ability students, the two-dimension format model was no better than the one-dimension model. Hypotheses one and two were thus supported. Based on the ANOVA results, there was a statistically significant interaction found between ability and format, indicating that the two groups did differ to some extent in dealing with different formats (MC vs. CR). However, hypothesis three may not be fully supported due to the small effect size.

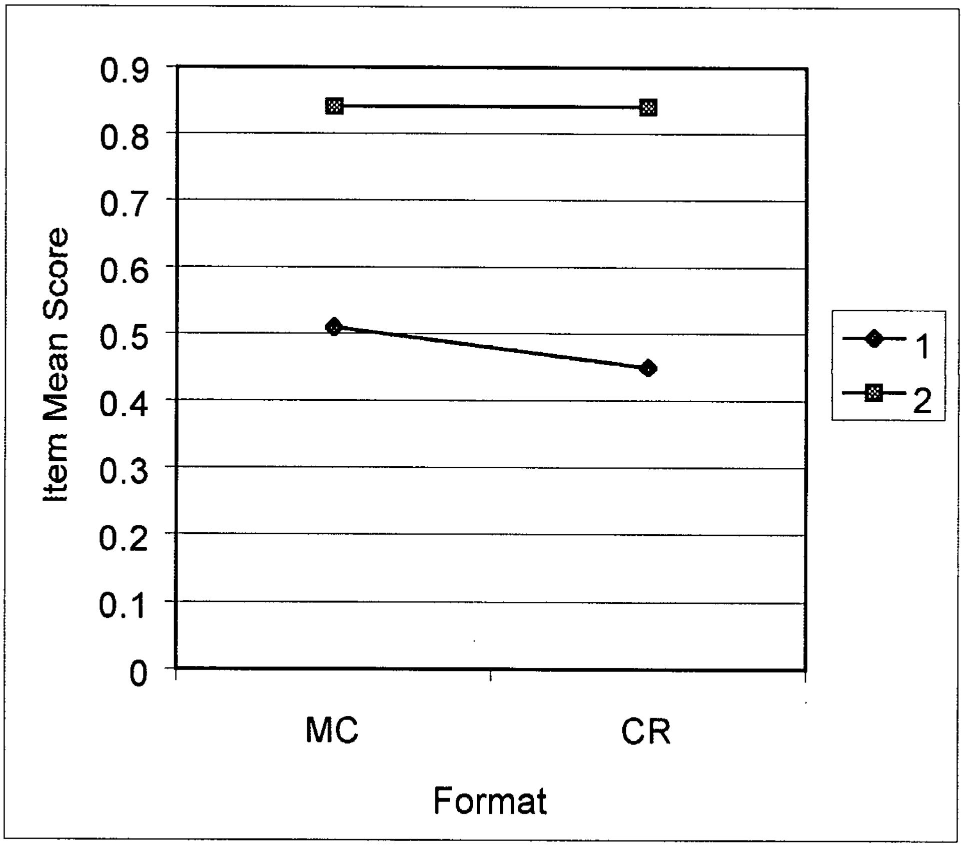


Figure 14. Comparison of High and Low Ability Students in terms of Format Differences (Data Set Three). MC – Multiple-Choice, CR – Constructed-Response, 1 – Low ability, 2 – High ability.

EXAMINATION OF DATA SET FOUR:
BRITISH COLUMBIA EXAMINATION (AUGUST 1998)

Missing Data

The analyses were based on the British Columbia Grade 12 Mathematics Examination (August 1998), from which 23 items were selected for the investigation. The sample (N = 1,429) was selected from the British Columbia Grade 12 Examination data bank. By looking at the missing rate in Table 4-34, it can be seen that only CR items were reported to have missing responses. CR item #6 had the largest missing response (15.4%). The average response rates for CR items were 93.2%. Speededness was not a problem because only one item at the end of the test had more than 10% of the missing responses.

Table 4-34

Percentage of Missing Data (Data Set Four)

Item	Valid Percent	Percent of Missing
1. MC1	100.0	0.0
2. MC2	100.0	0.0
3. MC3	100.0	0.0
4. MC4	100.0	0.0
5. MC5	100.0	0.0
6. MC6	100.0	0.0
7. MC7	100.0	0.0
8. MC8	100.0	0.0
9. MC9	100.0	0.0
10. MC10	100.0	0.0
11. MC11	100.0	0.0
12. MC12	100.0	0.0
13. MC13	100.0	0.0
14. MC14	100.0	0.0
15. MC15	100.0	0.0
16. MC16	100.0	0.0
17. MC17	100.0	0.0
Mean	100.0	0.0
18. CR1	94.9	5.1
19. CR2	96.2	3.8
20. CR3	**87.6**	**12.4**
21. CR4	98.6	1.4
22. CR5	97.1	2.9
23. CR6	**84.6**	**15.4**
Mean	93.2	6.8

Note. 23 items were selected from BC Grade 12 Mathematics Examination (August 1998). Sample Size is 1,429.

Constructed-Response Items

There were 6 CR items that were coded at multi-levels (maximum 4 levels). They were dichotomized and became dichotomous variables. For the four-point open-ended response items, scores of zero, one and two were re-coded to a value of zero, while scores of three and four were re-coded to a value of one. The information function was plotted to see if there was any information loss after dichotomizing the CR items. The information loss was 16% (see Figure 15, Appendix D). It seemed that the loss of information might affect the results from FIFA and MRCMLM to some extent. Therefore, the conclusions should be interpreted with caution.

Results from Classical Test Theory

From Table 4-35, it can be seen that students scored higher on MC items than on CR items (item mean score for MC items was 0.70; item mean score for CR items was 0.68). The item total correlation varied from 0.06 to 0.49. On average, CR items had higher item total correlation than MC items. The overall reliability (Cronbach's alpha) was 0.75.

Table 4-35

Descriptive Item Analysis[a] (Data Set Four)

Item	CL[b]	SC[c]	Item Mean	Std Dev	Item Total Correlation[d]
1. MC1	K	QR	0.90	0.30	0.28
2. MC2	U	QR	0.92	0.28	0.23
3. MC3	H	QR	0.42	0.49	0.33
4. MC4	U	TR	0.97	0.18	0.19
5. MC5	K	TR	0.64	0.48	0.38
6. MC6	H	TR	0.32	0.47	0.08
7. MC7	K	EL	0.91	0.28	0.31
8. MC8	U	EL	0.88	0.33	0.40
9. MC9	H	EL	0.47	0.50	0.20
10. MC10	K	PF	0.91	0.28	0.16
11. MC11	U	PF	0.77	0.42	0.35
12. MC12	H	PF	0.33	0.47	0.06
13. MC13	U	SS	0.97	0.17	0.19
14. MC14	H	SS	0.47	0.50	0.34
15. MC15	U	IC	0.87	0.33	0.19
16. MC16	K	IC	0.76	0.43	0.38
17. MC17	H	IC	0.47	0.50	0.34
18. CR1	U	SS	0.58	0.49	0.40
19. CR2	U	IC	0.69	0.46	0.42
20. CR3	U	IC	0.72	0.45	0.36
21. CR4	U	QR	0.70	0.46	0.49
22. CR5	U	EL	0.67	0.47	0.44
23. CR6	U	TR	0.72	0.45	0.29
Mean for MC			0.70		
Reliability[e]			0.62		
Mean for CR			0.68		
Reliability			0.65		
Mean			0.70		
Reliability for MC and CR			0.75		

Note. [a] 23 items from BC Grade 12 Mathematics Examination (August 1998, N=1,429). [b] CL-cognitive level, K-knowledge, U-understanding, H-higher mental process. [c] SC-sub-content, TR-trigonometry, QR-Quadratic Relations, EL-Exponential Logarithmic Functions, PF-Polynomial Functions, SS-Sequences and Series, IC-Introduction to Calculus. [d] Discrimination index-point-biserial. [e] Reliability-Cronbach's alpha.

RESEARCH QUESTION ONE

Exploratory Factor Analysis

Three models were tested to investigate whether item format factors affect the structure of mathematics test.

Hypothesis one: The test structure is unidimensional.

Hypothesis two: The test structure is two-dimensional.

Hypothesis three: The test structure is three-dimensional.

FIFA was applied as an exploratory approach to test the hypotheses. By looking at the variance and largest root indices in Table 4-36, it seemed that the one-factor model was the best of all the three models. The total variance accounted for by the model was 27.6%, and the largest root was 5.96. In terms of the two-factor model, the first factor was significant ($\lambda = 6.01$), but the second factor was not ($\lambda = 1.06$). Hypothesis two was rejected. Similarly, the first factor of the three-factor model was significant ($\lambda = 6.03$), but the second ($\lambda = 1.09$) and third factors ($\lambda = 0.56$) were not significant. Therefore, hypothesis three was not supported.

Examination of Factor Loadings

Based on the factor loadings in the one-factor solution (see Table 4-36), most MC and CR items loaded on the first factor with loadings greater than 0.30 except for three MC items (#6, #9, and #12). The examination of these three items revealed that they were not easy items (mean scores were 0.32, 0.47 and 0.33), and they had low item discrimination (0.08, 0.20 and 0.06).

The examination of the three items revealed that all of them required high level of conceptual understanding of the problem structure (e.g., equation). First, the examinees needed to conceptually understand the problem or equations provided. Second, schema knowledge of the computation rules was required to produce the correct answer. By looking at the other items that loaded highly on the factor (e.g., MC #8), it seemed that they required computation skills only. However, these 3 items required the examinees to have conceptual understanding of the mathematical expression in addition to computation skills. Therefore, these 3 MC items seemed to have high cognitive demand, which was consistent with the cognitive level defined by the TIMSS developers.

Table 4-36

Comparison of Factor Loadings of the Three Solutions Based on FIFA[a] (Data Four)

			One-factor[d]	Two- factor[e]		Three-factor		
Item	CL[b]	SC[c]	Factor 1	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3
1. MC1	K	QR	**0.576**	0.111	**0.534**[f]	0.263	**0.527**	-0.117
2. MC2	U	QR	**0.486**	**0.333**	0.185	**0.330**	0.155	0.068
3. MC3	H	QR	**0.494**	**0.327**	0.197	0.174	0.120	**0.309**
4. MC4	U	TR	**0.513**	0.243	**0.311**	**0.315**	**0.303**	-0.040
5. MC5	K	TR	**0.570**	**0.399**	0.202	0.290	0.136	0.248
6. MC6	H	TR	0.115	0.088	0.032	-0.091	-0.018	0.282
7. MC7	K	EL	**0.645**	0.087	**0.632**	0.165	**0.539**	0.080
8. MC8	U	EL	**0.751**	**0.506**	0.289	**0.568**	0.262	-0.004
9. MC9	H	EL	0.291	0.258	0.045	0.082	-0.001	0.282
10.MC10	K	PF	**0.335**	-0.120	**0.513**	-0.154	**0.401**	0.233
11.MC11	U	PF	**0.561**	**0.408**	0.183	**0.369**	0.126	0.147
12.MC12	H	PF	0.082	-0.005	0.100	-0.095	0.056	0.170
13.MC13	U	SS	**0.507**	**-0.437**	**1.043**	**-0.529**	**0.927**	**0.387**
14.MC14	H	SS	**0.488**	**0.332**	0.185	0.092	0.099	**0.434**
15.MC15	U	IC	**0.322**	0.198	0.145	0.039	0.065	**0.315**
16.MC16	K	IC	**0.601**	**0.379**	0.261	**0.384**	0.209	0.091
17.MC17	H	IC	**0.475**	**0.505**	-0.024	0.297	-0.079	**0.331**
18.CR1	U	SS	**0.580**	**0.563**	0.031	**0.366**	-0.023	**0.329**
19.CR2	U	IC	**0.629**	**0.751**	-0.119	**0.686**	-0.084	0.049
20.CR3	U	IC	**0.629**	**1.043**	**-0.507**	**0.959**	**-0.394**	-0.058
21.CR4	U	QR	**0.744**	**0.395**	**0.414**	**0.438**	**0.359**	0.072
22.CR5	U	EL	**0.667**	**0.619**	0.070	**0.595**	0.066	0.061
23.CR6	U	TR	**0.461**	**0.532**	-0.067	**0.435**	-0.068	0.129
Variance			27.64%	27.75%	3.70%	27.57%	4.16%	1.96%
Largest roots			5.96	6.01	1.06	6.03	1.09	0.56
Correlation								
Factor 1			-	1		1		
Factor 2				0.768	1	0.644	1	
Factor 3						0.621	0.397	1

Note. [a]23 items from BC Grade 12 Mathematics Examination (August 1998, N = 1,429). [b] CL-cognitive level, K-knowledge, U-understanding, H-high mental ability; [c] SC-sub-content, TR- trigonometry, QR-Quadratic Relations, EL-Exponential Logarithmic Functions, PF-Polynomial Functions, SS-Sequences and Series, IC-Introduction to Calculus. [d] Factor loadings for one-factor model are principal factor loadings. [e] Factor loadings for two- and three-factor models are promax factor loadings. [f] Bold numbers are loadings greater than the absolute value of 0.30.

MC #6. Determine **all** values of x for which the following expression is undefined,

$$\frac{\sin x}{1-\sec^2 x}, \text{ where } 0 \leq x < 2\pi$$

A. $\pi/2, 3\pi/2$ B. $0, \pi$ C. $0, \pi/2$ D. $0, \pi, \pi/2, 3\pi/2$

MC #8. Evaluate: $3 \log_7 2$ (Accurate to 2 decimal places).

A. 0.90 B. 0.92 C. 1.07 D. 1.13

MC #9. If $f(x) = ax$, determine **all** values of a such that $f^1(x) = f(x)$ for every x, where $f^1(x)$ is the inverse of $f(x)$.

A. -1 B. 1 C. ± 1 D. any non-zero real number.

MC #12 Determine the number of **rational** roots for the equation $x^5 - 2x - 1 = 0$.

A. 1 B. 2 C. 3 D. 5

The minor dimension in the two-factor solution seemed to be represented by four MC items (#1, #7, #10, and #13). The examination of the 4 items revealed that the problem structures were not very complex. Students were required to make the correct choice based on knowledge of rules or routine procedures. No complex procedures, reasoning, or explanation skills seemed to be needed. The item mean scores of the four items (0.90, 0.91, 0.91, and 0.97) also indicated that they were very easy items. Because the minor dimension was trivial based on the root and variance criteria, the one-factor model was the best model to be accepted. MC and CR items did seem to measure similar mathematical proficiency.

MC #1

Which of the following represents a parabola?

A. $xy = 1$ B. $x + y^2 = 1$ C. $x^2 + y^2 = 1$ D. $x^2 - y^2 = 1$

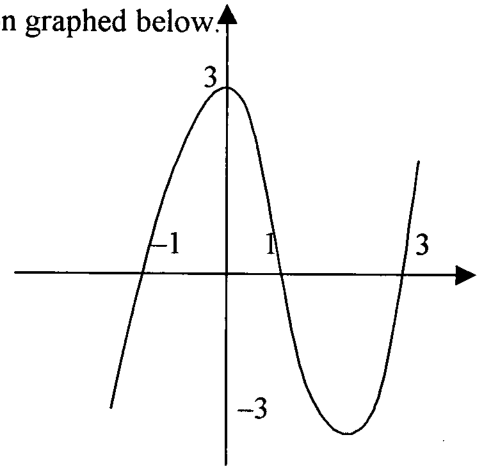
MC #7

Determine the exponential form of $\log_p r = m$.

A. $m = p^r$ B. $m = r^p$ C. $r = m^p$ D. $r = p^m$

MC#10

Determine the real zeros of the function graphed below.



A. 0, 2 B. −1, 1, 3 C. −1, 1, −3 D. −3, 3

MC #13

The *n*th term of a sequence is given by $t_n = \frac{2n}{n^2+1}$. Determine the 5th term.

A. $\frac{12}{37}$ B. $\frac{5}{13}$ C. $\frac{8}{17}$ D. 1.

The G^2 index (see 4-37) was examined to judge the significance of the dimensionality. The G^2 difference between the first and second factor was significant. However, the G^2 index may be affected by the large sample size (N = 1,429). Therefore, hypothesis one was supported based on the loading, variance, and root criteria.

Table 4-37

Change of the Likelihood Ratio G^2 in the Item Factor Analysis

Factor	G^2	df	p
2 vs. 1	74.74	22	0.000
3 vs. 2	36.92	21	0.020

Confirmatory Approach

The two-dimension format model was compared with the unidimensional model using MRCMLM. By looking at the deviance indices in Table 4-38, it seemed that the unidimensional model fit the data better than the two-dimension format model, $\chi^2_{(3)} = 2400.83$, $p < 0.001$. Based on the weighted fit indices, it seemed that both models did not fit the data very well. For the unidimensional model, there were 6 misfit items. However, for the two-dimension format model, 16 out of 23 items were misfit items, and 9 of them were serious misfit items. Therefore, it is possible to conclude that the unidimensional model was much better than the two-dimension format model.

Table 4-38

Comparison of Unidimensional and Two-dimension Format Models Based on MRCMLM[a] (Data Four)

					Unidimensional		Two-dimension	
Item	CL[b]	SC[c]	Estimate[d]	Error	MNSQ[e]	Wfit[f]	MNSQ	Wfit
1.MC1	K	QR	-2.538	0.092	0.96	-0.7	**1.51**	**26.3**[g]
2.MC2	U	QR	-2.749	0.099	1.01	0.2	0.83	**-2.6**
3.MC3	H	QR	0.363	0.058	0.98	-1.1	1.04	**2.1**
4.MC4	U	TR	-3.762	0.149	0.92	-0.6	0.81	-1.7
5. MC5	K	TR	-0.694	0.060	0.97	-1.2	0.91	**-5.1**
6. MC6	H	TR	0.903	0.062	**1.16**	**6.3**	**1.20**	**6.3**
7. MC7	K	EL	-2.700	0.098	0.91	-1.3	0.80	**-3.2**
8. MC8	U	EL	-2.281	0.085	0.87	**-2.4**	0.79	**-4.5**
9. MC9	H	EL	0.150	0.058	1.10	**4.7**	1.07	**4.2**
10.MC10	K	PF	-2.710	0.098	1.03	0.4	0.86	**-2.3**
11.MC11	U	PF	-1.400	0.067	0.98	-0.5	0.87	**-4.7**
12.MC12	H	PF	0.850	0.061	**1.17**	**6.9**	**1.21**	**6.8**
13.MC13	U	SS	-3.956	0.163	0.91	-0.6	0.79	-1.7
14.MC14	H	SS	0.163	0.058	1.01	0.7	0.99	-0.5
15.MC15	U	IC	-2.246	0.084	1.08	1.4	0.86	**-2.9**
16.MC16	K	IC	-1.378	0.067	0.96	-1.2	0.86	**-5.1**
17.MC17	H	IC	0.140	0.058	1.02	1.0	1.01	0.4
18.CR1	U	SS	-0.399	0.059	0.97	-1.2	0.96	**-2.1**
19.CR2	U	IC	-0.923	0.062	0.96	-1.4	0.97	-1.0
20.CR3	U	IC	-1.119	0.064	1.00	-0.1	0.99	-0.4
21.CR4	U	QR	-0.996	0.062	0.89	**-3.8**	**1.29**	**8.9**
22.CR5	U	EL	-0.844	0.061	0.92	**-3.0**	0.95	-2.0
23.CR6	U	TR	-1.087	0.063	1.07	**2.1**	0.92	**-2.6**
Deviance						31428.42		33829.25
Estimated[h] Parameters						25		28
Correlation (dimension)						-		0.829

Note. [a] 23 items from BC Grade 12 Mathematics Examination (August 1998, N = 1,429). [b] CL-cognitive level, K-knowledge, U-understanding, H-high mental ability; [c] SC-sub-content, TR- trigonometry, QR-Quadratic Relations, EL-Exponential Logarithmic Functions, PF-Polynomial Functions, SS-Sequences and Series, IC-Introduction to Calculus. [d] Estimate-difficulty parameter, [e] MNSQ-mean square residual. [f] Wfit-weighted fit. [g] Bold indicates misfit items. [h] the magnitudes are discussed in chapter 5.

Based on the latent distribution of the two-dimension format model (Figure 17), the first dimension (MC items) looked normally distributed while the second dimension (CR items) was slightly skewed positively. Students' distribution for the second dimension spread around more than the distribution of the first dimension.

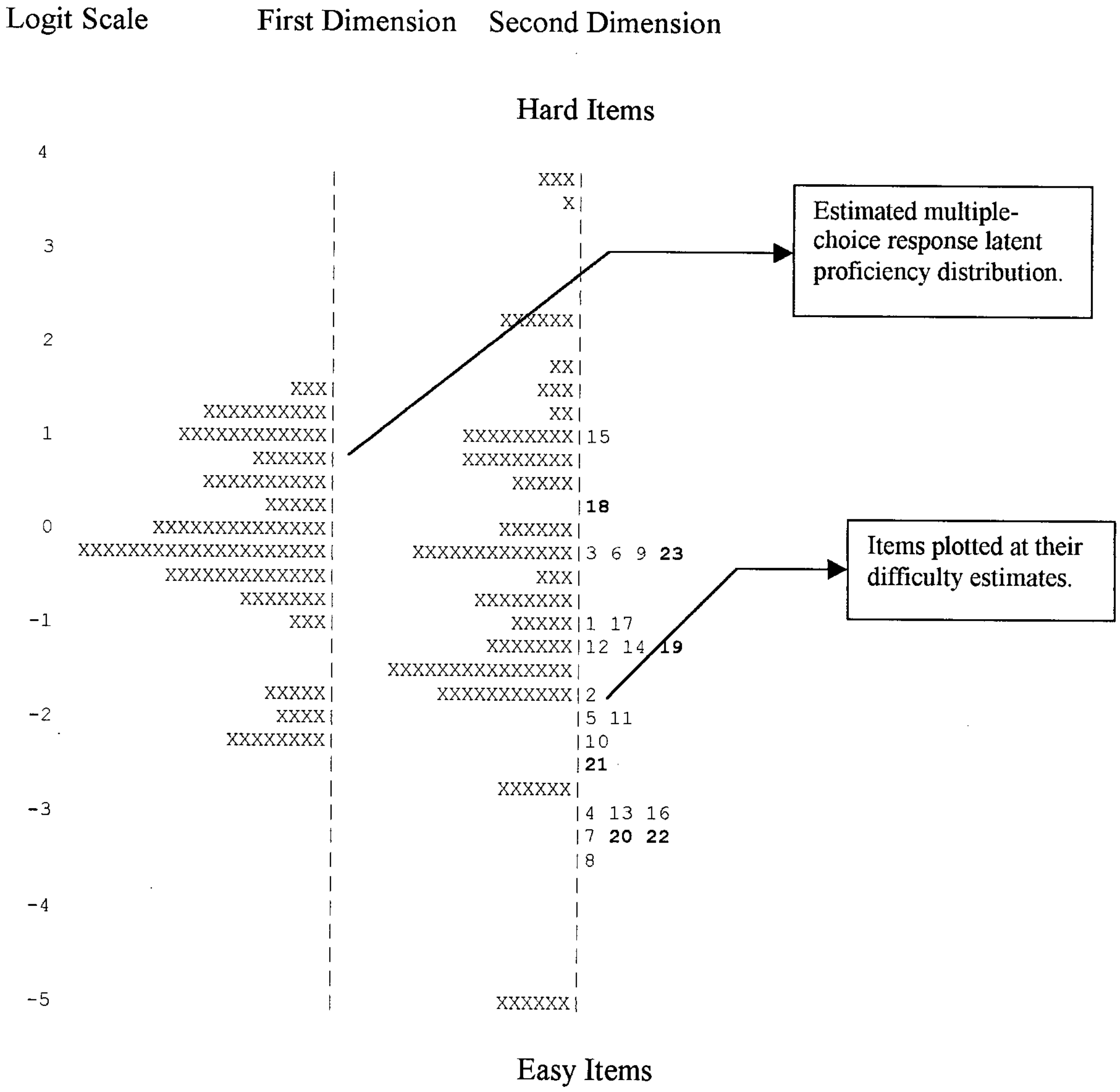


Figure 17. Map of Latent Distribution and Response Model Parameter Estimates for Two-dimension Format Model. Bold items are constructed-response items, non-bold items are multiple-choice items (23 items from BC Grade 12 Mathematics Examination, August, 1998).

Summary for Research Question One

The results from FIFA and MRCMLM indicated that the test structure was unidimensional. The evidence from FIFA revealed that the test structure was dominated by one factor—mathematical proficiency. The examination revealed that the 3 MC items requiring high cognitive ability did not load on the dominant factor, and the non-significant minor factor in the two-factor model was represented by 4 MC items requiring low cognitive ability. Therefore, the format differences did not seem to affect the dimensionality of the test. The results from MRCMLM also indicated that the unidimensional model was much better than the two-dimension format model. The hypothesis that MC and CR items represented two dimensions was not supported.

RESEARCH QUESTION TWO

Exploratory Factor Analysis

Three hypotheses were tested to investigate whether MC and CR items differ in measuring students' cognitive ability beyond knowledge level in mathematics.

Hypothesis One: The test structure is unidimensional.

Hypothesis Two: The test structure is two-dimensional.

Hypothesis Three: The test structure is three-dimensional.

FIFA was applied as an exploratory approach. Eighteen items designed to tap high cognitive levels (understanding and high mental process) in the British Columbia Grade12 Mathematics Examination were selected for the investigation. By looking at the variance and largest roots in Table 4-39, it seemed that the one-factor model was the best of all the three models. The total variance accounted for by the model was 26.2% and the largest root was 4.52, indicating that the first factor was a significant factor. Hypothesis one was supported. In terms of the two-factor model, the first factor was significant ($\lambda = 4.56$), but the second factor was not ($\lambda = 0.88$). Hypothesis two was rejected. Similarly, the first factor of the three-factor solution was significant ($\lambda = 4.59$), but the second ($\lambda = 0.88$) and third factors ($\lambda = 0.42$) were not significant. Therefore, hypothesis three was rejected.

Examination of Factor Loadings

All the loadings of the one-factor solution were above 0.30 except two MC items (#6 and #9). The cognitive analyses revealed that they required higher cognitive abilities.

The non-significant minor dimension in the two-factor solution was presented by 1 MC item #13. The cognitive analysis of this item revealed that it required simple computation skill. Because the minor dimension was trivial in terms of the root and variance criteria, the one-factor model was the best model to be accepted. MC and CR items beyond knowledge level seemed to measure the same mathematical proficiency.

Table 4-39

Comparison of Factor Loadings of the Three Solutions Based on FIFA[a]

			One-factor[d]	Two-factor[e]		Three-factor		
Items	CL[b]	SC[c]	Factor 1	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3
1.MC2	U	QR	**0.478**[f]	**0.451**	0.037	**0.433**	0.070	0.084
2.MC3	H	QR	**0.492**	**0.349**	0.204	0.272	0.137	**0.344**
3.MC4	U	TR	**0.498**	**0.365**	0.186	**0.323**	**0.358**	0.243
4.MC6	H	TR	0.128	-0.004	0.187	-0.122	0.111	**0.373**
5.MC8	U	EL	**0.736**	**0.647**	0.126	**0.741**	0.299	0.053
6.MC9	H	EL	**0.301**	0.199	0.146	0.111	0.141	**0.301**
7.MC11	U	PF	**0.549**	**0.530**	0.021	**0.512**	0.059	0.071
8.MC12	H	PF	0.069	-0.054	0.174	-0.105	0.089	0.261
9.MC13	U	SS	**0.426**	-0.255	**0.989**	**-0.317**	**1.094**	**1.213**
10.MC14	H	SS	**0.489**	**0.320**	0.242	0.165	0.195	**0.511**
11.MC15	U	IC	**0.308**	0.209	0.138	0.093	0.095	**0.329**
12.MC17	H	IC	**0.480**	**0.430**	0.071	**0.309**	-0.051	0.241
13.CR1	U	SS	**0.600**	**0.585**	0.018	**0.435**	-0.028	0.242
14.CR2	U	IC	**0.628**	**0.721**	-0.126	**0.666**	-0.147	-0.088
15.CR3	U	IC	**0.594**	**0.904**	**-0.413**	**0.776**	**-0.392**	**-0.320**
16.CR4	U	QR	**0.727**	**0.625**	0.140	**0.680**	**0.342**	0.162
17.CR5	U	EL	**0.660**	**0.701**	-0.059	**0.724**	0.072	-0.072
18.CR6	U	TR	**0.479**	**0.520**	-0.058	**0.428**	-0.058	0.068
Variance			26.22%	26.39%	4.19%	25.51%	4.82%	1.94%
Largest roots			4.52	4.56	0.88	4.59	0.88	0.42
Correlation								
Factor1				1		1		
Factor2			-	0.661	1	-0.185	1	
Factor3						0.636	-0.636	1

Note. [a] 23 items from BC Grade 12 Mathematics Examination (August 1998, N=1,429). [b] CL-cognitive level, U-understanding, H-high mental ability; [c] SC-sub-content, TR-trigonometry, QR-Quadratic Relations, EL-Exponential Logarithmic Functions, PF-Polynomial Functions, SS-Sequences and Series, IC-Introduction to Calculus; [d] Factor loadings for one-factor model are principal factor loadings. [e] Factor loadings for two- and three-factor models are promax factor loadings. [f] Bold numbers are loadings greater than the absolute value of 0.30.

Confirmatory Approach

The two-dimension between-item (format) model was compared with the unidimensional model by looking at the difference of the deviance indices (see Table 4-40). It seemed that the two-dimension format model fit the data better than the unidimensional model, $\chi^2_{(3)} = 13.59$, $p < 0.01$. The significant deviance difference between the models may be due to the large sample size. The correlation between the two dimensions was 0.90, indicating that the two dimensions measured similar ability.

The two-dimension within-item model and two-dimension between-item model seemed to work similarly because the deviance difference between the two models was only 2.71.

Therefore, the hypothesis that CR items might tap somewhat special characteristics other than the overall mathematical ability was not supported.

Table 4-40

Comparison of Unidimensional and Two-dimension Format Models Based on MRCMLM[a]

			Unidimensional	Two-dimension	Two-dimension
			MNSQ[d]	between-item	within-item
Item	CL[b]	SC[c]	(Wfit[e])	MNSQ (Wfit)	MNSQ(Wfit)
1.MC2	U	QR	0.98 (-0.3)	0.95 (-0.7)	0.97 (-0.5)
2.MC3	H	QR	0.99 (-0.1)	0.99 (-0.1)	0.99 (-0.2)
3.MC4	U	TR	0.97 (-1.5)	0.96 **(-2.2)**	0.96 **(-2.2)**
4.MC6	H	TR	0.93 (-0.5)	0.95 (-0.4)	0.96 (-0.2)
5.MC8	U	EL	1.01 (0.2)	0.98 (-0.7)	0.96 (-1.8)
6.MC9	H	EL	1.09 **(3.7)**	1.06 **(2.7)**	1.09 **(3.6)**
7.MC11	U	PF	0.92 (-1.1)	0.93 (-1.0)	0.94 (-0.8)
8.MC12	H	PF	0.93 (-1.2)	0.90 (-1.7)	0.94 (-1.1)
9.MC13	U	SS	1.05 **(2.7)**	1.04 (2.5)	1.06 **(3.5)**
10.MC14	H	SS	1.03 (0.4)	1.01 (0.1)	0.99 (-0.2)
11.MC15	U	IC	1.02 (0.6)	1.02 (0.5)	0.97 (-0.9)
12.MC17	H	IC	1.10 **(4.3)**	1.07 **(3.3)**	1.08 **(3.4)**
13.CR1	U	SS	0.93 (-0.5)	0.92 (-0.6)	0.99 (0.0)
14.CR2	U	IC	0.98 (-1.0)	0.99 (-0.3)	0.99 (-0.3)
15.CR3	U	IC	1.07 (1.2)	1.11 (1.7)	1.07 (1.3)
16.CR4	U	QR	0.95 (-1.3)	1.02 (0.6)	0.99 (-0.2)
17.CR5	U	EL	0.99 (-0.5)	1.00 (0.1)	1.02 (1.2)
18.CR6	U	TR	0.98 (-0.9)	0.99 (-0.5)	1.03 (1.3)
Deviance			23737.26	23723.67	23720.96
df			20	23	23
Correlation (dimension)			-	0.898	0.385

Note. [a] 18 items from BC Grade12 Mathematics Examination (August, 1998, N =1,429). [b]CL-cognitive level, U-understanding, H-higher mental process. [c] SC-sub-content, TR-trigonometry, QR-Quadratic Relations, EL-Exponential Logarithmic Functions, PF-Polynomial Functions, SS-Sequences and Series, IC-Introduction to Calculus. [d] MNSQ-mean square residual. [e] Wfit-weighted fit.

Summary for Research Question Two

The results from the exploratory factor analyses seemed to support the hypothesis that the test structure was unidimensional when selected items were measuring higher cognitive ability. However, the results from the confirmatory approaches indicated that the two-dimension between-item model was slightly better than the unidimensional model. The results from the confirmatory factor analysis should be interpreted with caution because it was possible that the deviance difference (13.59) was inflated due to the large sample size. The high correlation (0.90) between the two dimensions seemed to confirm that MC and CR items measured similar construct. In addition, the loadings of the exploratory factor analyses revealed that most MC and CR items loaded on the same factor. And the minor dimensions represented by several MC items in the two- and three-factor solution were not significant. The hypothesis that the test structure was unidimensional was thus supported, indicating that MC and CR items beyond knowledge level seemed to measure similar mathematical proficiency.

RESEARCH QUESTION THREE

In order to know whether high ability students differ from low ability students in dealing with different formats, the following hypotheses were tested:

Hypothesis One: MC and CR items are two dimensions for low ability students.

Hypothesis Two: MC and CR items are one dimension for high ability students.

Hypothesis Three: There is statistically significant interaction between format and ability level.

Hypothesis One

According to the differences between the deviance indices of the one- and two-dimension models in Table 4-41, the two-dimension format model seemed to fit the data better than the unidimensional model for the low ability students, $\chi^2_{(3)} = 29.45$, $p < 0.001$. The weighted fit indices of 18 items were within normal range. Therefore, hypothesis one was supported.

Table 4-41

Comparison of Unidimensional and Two-dimension Format Models on Low Ability Students' Responses[a]

					Unidimensional		Two-dimension	
Item[b]	CL[c]	SC[d]	Estimate[e]	Error	MNSQ[f]	Wfit[g]	MNSQ	Wfit
1.MC2	U	QR	-1.714	0.114	0.99	-0.2	0.97	-0.4
2.MC3	H	QR	1.396	0.103	1.05	0.9	1.03	0.4
3.MC4	U	TR	-2.645	0.163	0.96	-0.2	0.96	-0.2
4.MC6	H	TR	1.112	0.096	1.03	0.7	1.00	0.1
5.MC8	U	EL	-1.021	0.094	0.96	-1.0	1.00	-0.0
6.MC9	H	EL	0.724	0.089	1.04	1.3	1.04	1.3
7.MC11	U	PF	-0.284	0.084	0.98	-1.0	1.01	0.3
8.MC12	H	PF	1.057	0.094	1.04	0.8	1.02	0.5
9.MC13	U	SS	-2.846	0.177	0.96	-0.2	0.96	-0.2
10.MC14	H	SS	1.084	0.095	1.04	0.8	1.02	0.4
11.MC15	U	IC	-1.335	0.102	0.98	-0.3	0.97	-0.4
12.MC17	H	IC	1.167	0.097	1.03	0.6	1.02	0.4
13.CR1	U	SS	0.787	0.090	1.02	0.6	1.00	0.0
14.CR2	U	IC	0.379	0.085	1.02	0.9	0.99	-0.4
15.CR3	U	IC	0.068	0.084	1.01	0.7	0.99	-0.4
16.CR4	U	QR	0.322	0.085	1.00	-0.0	0.99	-0.4
17.CR5	U	EL	0.358	0.085	0.99	-0.6	1.00	0.2
18.CR6	U	TR	-0.107	0.084	1.01	0.9	0.99	-0.5
Deviance						12004.46		11975.01
df						20		23
Correlation (dimension)						-		-0.128

Note. [a] Low ability - students at the low 40% of the total score (N = 589); [b] 18 items from BC Grade12 Mathematics Examination (August 1998); [c] CL-cognitive level, U–understanding, H-higher mental process; [d] SC-sub-content, TR-trigonometry, QR-Quadratic Relations, EL-Exponential Logarithmic Functions, PF-Polynomial Functions, SS-Sequences and Series, IC-Introduction to Calculus. [e] Estimate-difficulty parameter, [f] MNSQ-mean square residual. [f] Wfit-weighted fit.

Hypothesis Two

According to the differences of the deviance indices between the unidimensional and two-dimension models in Table 4-42, the two-dimension format model seemed to fit the data no better than the unidimensional model for the high ability students, $\chi^2_{(3)} = 5.84$, $p > 0.05$. Hypothesis two was thus supported.

Table 4-42

Comparison of Unidimensional and Two-dimension Format Models on High Ability Students' Responses[a]

					Unidimensional		Two-dimension	
Item[b]	CL[c]	SC[d]	Estimate[e]	Error	MNSQ[f]	Wfit[g]	MNSQ	Wfit
1.MC2	QR	U	-3.775	0.246	0.96	-0.1	0.99	0.0
2.MC3	QR	H	-0.536	0.079	1.03	1.2	1.00	-0.1
3.MC4	TR	U	-5.018	0.449	0.96	0.1	1.00	0.1
4.MC6	TR	H	0.506	0.079	1.00	-0.0	0.99	-0.5
5.MC8	EL	U	-4.424	0.336	0.96	-0.0	1.00	0.1
6.MC9	EL	H	-0.450	0.078	1.03	1.5	1.00	0.2
7.MC11	PF	U	-2.498	0.139	0.98	-0.1	1.01	0.1
8.MC12	PF	H	0.463	0.078	1.00	0.2	1.03	1.3
9.MC13	SS	U	-5.532	0.579	0.96	0.1	0.99	0.2
10.MC14	SS	H	-0.663	0.080	1.01	0.4	1.02	0.7
11.MC15	IC	U	-2.964	0.169	0.97	-0.2	1.00	0.1
12.MC17	IC	H	-0.767	0.082	1.00	0.1	1.01	0.5
13.CR1	SS	U	-1.540	0.098	1.00	-0.0	1.06	1.0
14.CR2	IC	U	-2.177	0.122	1.00	0.0	1.01	0.2
15.CR3	IC	U	-2.319	0.129	0.99	-0.0	1.03	0.3
16.CR4	QR	U	-2.352	0.131	0.97	-0.3	1.01	0.1
17.CR5	EL	U	-2.133	0.120	0.99	-0.1	1.00	0.1
18.CR6	TR	U	-1.919	0.111	0.98	-0.3	1.02	0.3
Deviance						9721.15		9715.31
df						20		23
Correlation (dimension)						-		0.148

Note. [a] High ability - students at the high 40% of the total score (N = 709); [b] 18 items from BC Grade12 Mathematics Examination (August 1998); [c]CL-cognitive level, U-understanding, HM-higher mental process; [d]SC-sub-content, TR-trigonometry, QR-Quadratic Relations, EL-Exponential Logarithmic Functions; PF-Polynomial Functions, SS-Sequences and Series, IC-Introduction to Calculus. [e] Estimate-difficulty parameter, [f] MNSQ-mean square residual. [g] Wfit-weighted fit.

Hypothesis Three

Descriptive statistics as well as the ANOVA analyses were used to see how high ability students differ from low ability students in terms of their performance on the different formats. The item mean scores in Table 4-43 showed that the high ability group performed better on both formats than the low ability group. Low ability students performed better on MC items than on CR items; however, high ability students performed better on CR items than on MC items.

Table 4-43

Mean Scores of High and Low Ability Students' Responses on MC and CR Items

	Ability	Item Mean	Std. Deviation	N
Multiple-choice	low	0.5301	0.1227	589
	high	0.7627	0.1072	709
	Total	0.6572	0.1628	1298
Constructed-response	low	0.4298	0.2225	589
	high	0.8796	0.1407	709
	Total	0.6755	0.2888	1298

Note. Low ability - students at the low 40% of the total score (N = 589); High ability – students at the high 40% of the total score (N = 709).

In Table 4-44, the statistics showed that high and low ability students differed on the item mean scores of the two formats, $F_{(1, 1296)} = 323.334$, $p < 0.001$ (see also Figure 18). The effect size f of the interaction was regarded as large (Cohen, 1988, p.287). Therefore, the hypothesis that high ability students differed from low ability students in dealing with MC and CR items was supported.

Table 4-44

Analysis of Variance for High and Low Ability Students' Mean Scores on MC and CR Items

Source	SS	df	MS	F	Eta (η^2)	f
Between			Between			
Group	74.904	1	74.904	3263.20*	0.716	1.588
Error (Group)	29.749	1296	0.023			
Within			Within			
Format[a]	0.045	1	0.045	1.902	0.001	0.000
Format * Group[b]	7.593	1	7.593	324.334*	0.200	0.500
Error (Format)	30.341	1296	0.023			

Note. [a] Format-MC and CR; [b] Low ability - students at the low 40% of the total score (N = 589); High ability - students at the high 40% of the total score (N = 709). * p < .001.

Summary for Research Question Three

The results from the confirmatory approaches indicated that, for low ability students, the two-dimension format model was a better model than the unidimensional model. However, for high ability students, the two-dimension format model was no better than the unidimensional model, indicating that the two different formats measured similar constructs. Hypothesis one and two were thus supported. Based on the results from ANOVA, the interaction between ability and format was statistically significant, indicating that the two groups differed in dealing with the different formats (MC vs. CR). Hypothesis three was thus supported.

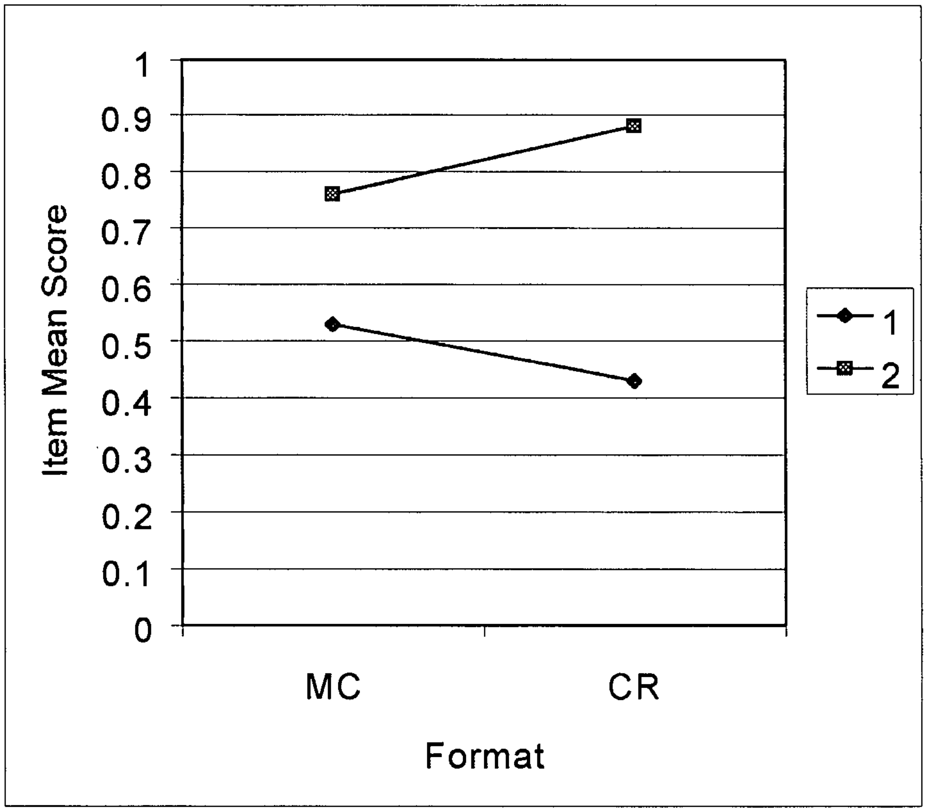


Figure 18. Comparison of High and Low ability Students in terms of Format Differences (Data Set Four). MC – Multiple-Choice, CR – Constructed-Response, 1 – Low ability, 2 – High ability.

CHAPTER SUMMARY

The current study addressed three research questions by analyzing four large-scale mathematics assessment data sets (TIMSS Grade 3 and Grade 4; TIMSS Grade 7 and Grade 8; British Columbia Grade12 April Examination; British Columbia Grade12 August Examination):

1. How do item formats affect the dimensionality of the test structure?
2. How do MC and CR items differ in measuring students' higher cognitive ability in the mathematics examinations?
3. How do differences in item formats affect students' performance at different levels?

In this chapter, the results of the analyses of the four data sets were presented. In terms of the first research question, the unidimensionality hypothesis was supported by both exploratory (FIFA) and confirmatory factor analysis (MRCMLM, see Table 4-45). All the data structures seemed to be mainly dominated by one factor — mathematical ability (proficiency). The non-significant minor dimensions found in the exploratory factor analyses of data set one and two seemed to be due to the existence of the item local dependence and low level of cognitive ability (computation skills). For data set three, the non-significant minor dimension was represented by several CR items that had high cognitive demand. For data set four, the non-significant minor dimension was represented by several MC items that had low cognitive demand. In addition, the examination of the factor loadings revealed that those items that failed to load with other items on the dominant factor were those that required different cognitive demand. Therefore, the differences between MC and CR items did not affect the dimensionality of the test structures, and MC and CR items seemed to measure similar mathematical proficiency.

In terms of the second research question, the unidimensional model was supported in the exploratory factor analysis across the four analyses. The results from exploratory factor analysis indicated that the two-factor model was no better than the unidimensional model in representing the test structures based on the loading, variance, and root criteria. However, the two-dimension format model was supported by the confirmatory factor analyses across the four analyses. The results showed that the difference of the deviance between the unidimensional and two-dimension between-item (format) model (MC vs. CR) was statistically significant. Such results contradicted the findings from FIFA. The statistical significance of the confirmatory factor analysis should be interpreted with caution because the large sample size may inflate the deviance difference (Bock et al., 1988). The high correlation between the two dimensions (MC vs. CR) in all four data sets indicate that the two formats assessed similar constructs. In addition, the fact that the two-dimension within-item model was no better than the two-dimension between-item format model indicated that CR items did not differ much from the MC items in measuring the mathematical ability. Further, for data sets one and two, the examination of the factor loadings revealed that the non-significant minor dimensions appeared to be due to the existence of the local dependence of the two pairs of CR items. For data sets three and four, the non-significant minor dimensions were due to several items that had either high or low cognitive demand. Therefore, it seemed that format differences were not the reason for the non-significant minor dimensions. Instead, cognitive complexity seemed to be the reasons for the minor dimensions. It can be concluded that MC and CR items designed to tap similar cognitive ability beyond knowledge level did not differ in measuring students' mathematical proficiency.

Table 4-45

Comparison of the Unidimensional and Two-dimension Format Models Based on MRCMLM

	Unidimensional	Two-dimension	Deviance	
		Between-item	Difference	p
	Deviance (df)	Deviance (df)	M 2[a] vs.1[b] (df)	
Data Set One	68282.62	70960.19	2677.57	
(N=2011)	31	34	3	0.000
Data Set Two	60591.30	63036.32	2445.02	
(N=2073)	28	31	3	0.000
Data Set Three	35927.44	38187.14	2259.70	
(N=1718)	25	28	3	0.000
Data Set Four	31428.42	33829.25	2400.83	
(N=1430)	25	28	3	0.000

Note. [a] Two-dimension between-item format model. [b] Unidimensional model.

In terms of the third research question, the hypothesis that high and low ability students differed in dealing with MC and CR items was supported. For data sets one, three, and four, the data structures were unidimensional for high ability students and two-dimensional for low ability students. For data set two, the data structure was unidimensional for low ability students and two-dimensional for high ability students. The reason for the above finding might be that the CR items of data set two were more difficult than those CR items in the other three data sets. Therefore, high ability students might be able to use different strategies to come up with the answers while low ability students failed on those difficult items no matter how hard they tried. Generally speaking, for high and low ability students, the data structure was different in all the four analyses. In addition, there was a statistically significant interaction found between low and high ability students in terms of the item mean scores according to the ANOVA results (see Table 4-46).

Table 4-46

Statistics of Format and Ability Interaction for the Four Data Structures Based on ANOVA Analysis

	F	Hypothesis df	Error df	Sig.	Eta Square	f
Data Set One	208.199	1	1262	0.000	0.142	0.407
Data Set Two	139.428	1	1346	0.000	0.094	0.322
Data Set Three	27.010	1	1562	0.000	0.017	0.132
Data Set Four	324.334	1	1296	0.000	0.200	0.500

The results from the four analyses supported the hypothesis that the format differences did not affect the unidimensionality of the test structures. It seemed that item local dependence and different level of cognitive demand were the main reasons for the existence of the non-significant minor dimensions. The examination of factor loadings and cognitive demand revealed that the items that failed to load with other items on the dominant factor required different cognitive demands. MC and CR items did not differ when they were supposed to measure mathematical proficiency beyond knowledge level. High and low ability students performed differently in dealing with the different item types. It appeared that low ability students performed better on MC items than on CR items across all the data sets. However, for high ability students, this was not the case. For data sets one, three and four, high ability students did similarly or better on CR items than on MC items. For data set two, high ability students did better on MC items than CR items.

CHAPTER V: SUMMARY AND DISCUSSION

SUMMARY

The purpose of the study was to determine: (1) whether differences between MC and CR item types lead to multidimensionality in mathematics assessment; (2) whether MC and CR items differ in the degree to which they assess cognitive ability beyond knowledge levels based on Bloom's learning taxonomy; and (3) whether high and low ability students differ in their performances on MC and CR items. Three main research questions and related hypotheses were tested based on four different large-scale assessments in mathematics. The results indicated that the four test structures appeared to be unidimensional. Further, MC and CR items did not differ in assessing students' cognitive ability beyond knowledge level. High and low ability students differed to some extent in dealing with MC and CR item types.

Research Question One

In order to assess whether the differences between MC and CR item types lead to the multidimensionality of the test structures, both exploratory and confirmatory factor analyses were applied using the Full-information Factor Analysis (FIFA) and the Multidimensional Multinomial Random Coefficients Logit Model (MRCMLM). The analyses of the four examinations using the exploratory approach indicated that all the test structures were unidimensional because the one-factor model fit the data better than the two- and three-factor models. The examinations of the factor loadings revealed that those items that failed to load with other items on the dominant factor were those that required either lower (e.g., no computation skill) or higher cognitive demand (e.g., multiple procedures). The non-significant minor dimensions found in the exploratory factor analyses of data set one and two seemed to be due to the existence of the item local dependence and low cognitive demand (no computation skills). In data set three, the non-significant minor dimension was represented by several CR items that had high cognitive demand. In data set four, the non-significant minor dimension was represented by several MC items that had low cognitive demand. The evidence from the four analyses in the present investigation consistently revealed that the one-factor model was better than the two-factor model (MC vs. CR).

Research Question Two

The items that were designed to tap higher cognitive ability beyond knowledge level were selected in order to assess whether MC and CR items differed. The high cognitive levels beyond knowledge acquisition included complex procedures, understanding, and higher mental process as defined by the test developers. In the four analyses of the present investigation, the results from the exploratory factor analysis (FIFA) indicated that the test structure was unidimensional. The unidimensional model seemed to fit the data better than the two- and three-factor models when the test items were designed to assess ability beyond knowledge acquisition.

However, the evidence from the confirmatory approach (MRCMLM) appeared to indicate that the two-dimension between-item (format) model fit the data better than the unidimensional model according to the difference of the deviance between the two models. Because of the large sample sizes used in the four analyses (2,011, 2,073, 1,780, and 1,429), it is possible that the difference between the deviance of the two models was inflated. If smaller sample size was used in the study, the difference of the deviance may have become smaller. The correlation between the MC and CR items ranged from 0.86 to 0.91 in the four data sets, indicating that they measured similar constructs. The two-dimension within-item model in each data set turned out to be no better than the two-dimension between-item model except in data set one, which was the further evidence that CR items did not tap different construct other than the general mathematical proficiency. In addition, in data set one and two, the examination of factor loadings revealed that the non-significant minor dimensions seemed to be due to the existence of the local dependence between two pairs of CR items and low cognitive demand. In data set three, the non-significant minor dimensions were due to several items that had high cognitive demand (multiple procedures). In data set four, the minor dimension was represented by the items of low cognitive demand (simple computation skill). Generally speaking, it seemed that the items that did not load on the factor in the one-factor solution were those that had either lower or higher cognitive demand, or local dependent items. Therefore, local dependent items and different levels of cognitive demand instead of format differences were the reason for the non-significant minor dimensions. In terms of the second research question, it can be concluded that MC and CR items designed to tap higher cognitive ability beyond knowledge level measured similar ability in mathematics.

Research Question Three

In each study, high and low ability students were compared to see if they dealt with MC and CR items differently. The same set of items can result in unidimensional data for one group of examinees and multidimensional data for the other group because the dimensionality is a function of the interaction between examinees and the items (Ackerman, Simpson, & de la Torre, 2000). The results from the confirmatory approaches indicated that the test structure was different for low and high ability students. In the analyses of data set one, three, and four, for low ability students, the two-dimension between-item (format) model was the better model. It appeared that the two formats were two somewhat distinct constructs. However, for high ability students, the two-dimension format model was no better than the unidimensional model. It was very interesting to note that the above situation was not the case in data set two. The data structure was two-dimensional for high ability students and unidimensional for low ability students. One reason for this finding might be that the CR items in data set two were relatively difficult compared with those in the other tests. It was possible that low ability students frequently failed on those difficult items regardless of item type, whereas high ability students were successful at those difficult items because they used different strategies when encountering different item types. Therefore, it was not difficult to explain the findings in other easy examinations, where high ability students could deal with items using similar strategies and low ability students were able to get correct answers by using different strategies.

Based on the results from the analysis of variance, there was a statistically significant interaction between ability and format on the item mean scores in all four analyses. Only one small effect size (0.132) for the interaction was found in the four analyses (see Table 4-46), indicating that the two groups differed in their performances of different formats (MC vs. CR). The hypotheses that high and low ability students differed in dealing with different item types were thus tenable.

DISCUSSION

In response to the first research question, the results from the four analyses in the present investigation indicated that the unidimensional model was the best model compared with the two-dimension format model (MC vs. CR). If different item types created two different dimensions and the correlation between them was low, then it was obvious that the unidimensionality assumption of IRT would be violated. Using the unidimensional IRT model to scale students cannot provide accurate information. The results of the four assessments did support the conclusion that MC and CR item did not differ in assessing students' mathematical ability. Such finding was consistent with what was found in some previous studies (Traub & Fisher, 1977; Bennett et al., 1991; Ercikan et al., 1998). Ercikan et al. (1998) concluded that MC and CR items can be calibrated to create a common scale using unidimensional IRT models in four content areas including mathematics. The multidimensional IRT models (FIFA and MRCMLM) applied in the present investigation appeared to confirm the results from the previous studies based on linear factor analysis and the unidimensional IRT model.

However, the conclusions from the present study contradicted what was found in some other studies (O'Neil & Brown, 1998; Walker & Beretvas, 2000). O'Neil and Brown (1998) investigated the effect of item format on metacognitive and affective processes of children using metacognition and affect questionnaires. The results of their study indicated that CR and MC items had differential effects. CR items included more cognitive strategy usage, less self-checking, and greater worry than did MC items. The examination of the factor loadings in the present study appeared to reveal that the main reasons for the non-significant minor dimensions involved the following two situations: (1) the item local dependence due to the same context; (2) items of different cognitive demand. From the examination of the items, it can be seen that the fact that some items did not load with other items on the main factor was due to the low cognitive demand (e.g., no computation skills required). Some items requiring higher cognitive ability (e.g., multiple procedures) were the reasons for the existence of the non-significant minor dimensions. It did not seem to be the case that all CR items measured higher ability than MC items did. Most MC and CR items loaded together on the same factor in the one-factor solution. The high correlation between MC and CR items also suggested that they measured similar constructs. All the evidence in the four analyses appeared to indicate that MC and CR items could be scaled together.

However, according to Bennett (1993), different item formats may produce highly correlated scores even when distinct processes are involved. He further pointed out that such highly correlated scores might be treated equivalently for some purposes, but they may not be measures of the same attribute. Therefore, the examination of the cognitive demand of items should be used to assist the interpretation of the statistical results. In the present investigation, MC and CR items appeared to measure similar mathematical proficiency because most MC and CR items loaded together on the dominant factor. The examination of the cognitive demand showed that the items with low loadings on the dominant factor or items with high loadings on the minor dimension required different cognitive demand. It did not seem to be the case that all the CR items measured higher cognitive ability than did the MC items in the four analyses. Some CR items seemed to require simple answers, whereas some MC items were very challenging.

More emphasis in mathematics examination nowadays is placed on problem solving, mathematical reasoning, and communication (graphic, numerical, and writing) ability. It is possible that some mathematics examinations may become multidimensional. According to Walker and Beretvas (2000), the mathematics examination including both MC and CR items in their study was multidimensional: one dimension (CR items) representing an examinee's ability to communicate about mathematics and the other dimension (MC items) representing an examinee's ability to solve mathematical problems. Although such findings were not supported in the present investigation, it is possible that the degree of format differences might vary from assessment to assessment. Therefore, the violation of the unidimensionality assumption might be a very serious problem for some assessments, but not for others.

Fortunately, the latest development in IRT has been extended from unidimensional to multidimensional IRT models (Ackerman, 1992; Reckase, 1997). Thus, it is important to check whether the test structure is unidimensional or multidimensional before the decision is made as to whether a unidimensional or a multidimensional IRT model should be used for calibration and reporting purposes.

In response to the second research question, the results of the present investigation supported Hancock's findings (1996). Most previous researchers, except Hancock (1996), did not incorporate cognitive framework into their analyses. In his study, Hancock incorporated Bloom's taxonomy as the framework to investigate the format differences within each taxonomic level. Comparison of format differences without incorporating such a framework was difficult to interpret. In an approach similar to Hancock's, the items of higher cognitive level were selected in order to address the second research question in the present investigation, and MC and CR item types were compared at similar cognitive levels.

Due to the large sample size used in the four analyses, the weighted fit statistics and the significance of the deviance difference based on MRCMLM might be inflated (see Table 4-5, 4-16, 4-27, 4-38). Especially, the deviance for the two-parameter model was much larger than that of the more constrained one-parameter model. This could be due to the number of nodes, the number of iterations, or the fact that the two-dimension confirmatory model was a bad model for the data. The third reason is most likely because the number of nodes and the number of iterations were set to maximum allowable in the software.

In the study, evidence such as the variance explained, root criteria, and correlation were examined in order to determine the dimensionality. Additionally, the examination of the factor loadings revealed that the non-significant minor dimensions were due to the existence of item local dependence and different cognitive demand. Although all the MC and CR items were supposed to tap similar cognitive ability designed by the test developers, several items loading on the non-significant minor dimensions were found to be different somehow from the other items in terms of the different cognitive demand required (e.g., no computation skill involved). Generally speaking, most MC and CR items loaded together on the same factor in the one-factor solution, and MC and CR items correlated highly. Therefore, MC and CR items can be considered to measure similar mathematical proficiency.

Although MC and CR items beyond knowledge level appeared to assess similar latent abilities according to the statistical results, the cognitive demand in answering MC and CR items might be different. According to Snow (1993, p.45), when an answer is to be chosen from a list, all that is necessary is to search for a match with information in memory and check the answer if a match is found; some knowledge and problem-solving skills as well as evaluation ability are needed. However, when the answers must be produced by the examinee, he or she has to search for the information in memory to a larger extent, reason more logically, and finally evaluate the answer without any hint. Therefore, in some situations, it is possible that CR and MC items may put different cognitive demands on examinees, even though the test structure incorporating both MC and CR item types is unidimensional.

High and low ability students differed in dealing with different item formats. The analyses of the four different assessments in mathematics indicated that high ability students handled CR item types better than the low ability students. It seemed that students of different ability differed in how they dealt with different item formats. The results from MRCMLM showed that the data structure was unidimensional for high ability students and two-dimensional for low ability students in the analyses of one, three, and four. Students of low ability may have applied certain strategies to answer the item correctly by looking at the distractors of the MC items or by guessing based on partial knowledge. Such strategies might not work for CR items. On the other hand, students of high ability may have applied similar strategies to handle both formats because they had organized knowledge structures. Therefore, it is possible that high ability students may often apply a simple working forward strategy, whereas low ability students may often apply more powerful strategy such as means ends method (Chi, Glaser & Rees, 1982).

However, the results from data set two indicated that the data structure was unidimensional for low ability students and two-dimensional for high ability students. It seemed that this conclusion supported the hypothesis by Snow and Lohman (1993), who claimed that the data structure might be unidimensional for low ability students and multidimensional for high ability students. Further studies seem to be necessary to help to find out why the conclusion from data set two was different from that of the other three data sets. One possible reason for the different results might be due to the difficulty level of the test items. The item mean scores on CR items for high and low ability students were 0.35 (data set two), which was low compared with that of the other three data sets. The mean scores of CR items for data set one, three, and four were 0.48, 0.63, and 0.68 respectively. Snow's (1993) observation that students shifted strategies when they dealt with items of different difficulty levels seemed to support the above interpretation. Snow (1993) described the situation as follows:

> Ability differences in strategic processing are marked. An able student might turn a MC item into a CR item by first processing the stem to construct a possible answer, and only then scanning the response alternatives to find a match. A less able student might conduct a feature comparison search between stem and alternatives from the start in order to eliminate response alternatives, with no attempt at mental construction. And ability is relative to item difficulty; the able student may also shift to a response elimination strategy on difficult items, and the less able student may use mental construction on an item that is easy enough to allow it (p.57).

However, some findings in the previous studies by cognitive psychologists seemed to be contradictory when high and low ability students were compared. Some have claimed that low ability students often apply different strategies to deal with test items, whereas high ability students generally use a simple working forward method (Chi et al., 1982; Gagne, 1985). Other researchers believe that a test might be unidimensional for low ability students because all problems are relatively new for them, and thus require the same general problem-solving skills, whereas high ability students may show different patterns of skill development on different types of problems (Snow & Lohman, 1993).

Although many tests are unidimensional and one score might be enough for the reporting purpose, it seems that using a profile of scores instead of one scale score might be better articulated and more precise for diagnostic decisions, and for licensure decisions. For example, if each student has two scores on a mathematics test (MC and CR scores), it will be easier for teachers to identify those students who obtain low scores on the CR items. Those students may need extra help from teachers and parents or need extra motivation to complete those CR items.

IMPLICATIONS

Implications for Practice of Mathematics Assessment

The results of the study will be very useful to educators, researchers, and assessment specialists in determining whether to use single item type or both MC and CR items to assess students' mathematical ability. It appears that MC and CR formats do not lead to the multidimensionality of the test structures in mathematics. Therefore, unidimensional IRT model can be used to calibrate a mathematics test incorporating both MC and CR items, which support the conclusions from the previous researchers (Ercikan, et al., 1998).

However, the test structures of the four examinations appeared to be two-dimensional (MC vs. CR) for the subgroups (e.g., high and low ability students in the present study). A certain item type might be very difficult to certain group of students. Therefore, it is possible that the data structure can be multidimensional in some situations. A multidimensional IRT model is better than the unidimensional model to apply when the unidimensionality assumption is violated in the data set. Reporting scores for each dimension might be better than reporting one overall score, which might help teachers identify students' weakness in certain area, and find ways to help them to improve their performances. Generally speaking, researchers should make efforts to apply the appropriate models to analyze the data and produce accurate reports on students' performance.

Implications for Methodology of Assessing Dimensionality

Both exploratory (FIFA) and confirmatory approaches (MRCMLM) were used to address the research questions. These two recently developed multidimensional IRT models provided the possibility of exploring and confirming the test structures in a much better and more flexible way. In the previous research studies, no researchers applied the above two models together to investigate the dimensionality. In addition, the analyses incorporating the cognitive demand helped to explain the nature of the test structures (i.e., unexpected loadings in the exploratory factor analyses). Few researchers applied cognitive models along with the psychometric models to investigate the differences between item types. The multiple methods applied in the present study certainly help to provide stronger evidence regarding the research questions.

Implications for Theory Development

Bloom's learning taxonomy was used as a framework to address the second research question. Items tapping higher cognitive levels were selected for the investigation. The question of whether MC and CR items differed in measuring students' higher cognitive ability was answered. The findings also indicated that item local dependence and cognitive demand instead of format differences were the reasons for the minor dimensions. In addition, the evidence that high and low ability students differed in dealing with MC and CR items supported Snow's (1993) idea regarding the differences between able and less able students. According to the present investigation, it appeared that the test structure incorporating both MC and CR items was unidimensional for high ability students and two-dimensional (MC vs. CR items) for low ability students when the test was relatively easy. On the other hand, when the test was relatively difficult, it became unidimensional for low ability students and two-dimensional (MC vs. CR items) for high ability students.

LIMITATIONS OF THE STUDY

There were two primary limitations of the study. First, the psychometric models used in the study had limitations. The FIFA model can handle only dichotomous variables, so some information was lost when polytomous items were dichotomized. The total test information loss was trivial in data sets 1 and 2, because only 3 items were coded as 0, 1, and 2, whereas other CR items were coded as 0 and 1. More information loss occurred in data sets 3 and 4 because all the

CR items were coded beyond 3 levels. Although the dichotomization of the polytomous items was based on the distribution of students' responses, such procedure was a limitation of the study because the results may be affected to some extent by the loss of information.

The second limitation of the study was related to the generalizability of the study. Although the four mathematics examinations varied across different purposes, levels, and time points, the conclusions relating to the comparison of different formats in mathematics examinations cannot be generalized to other subject areas (e.g., science, social studies, and language). The previous studies suggested that MC and CR items differed in the writing domain. However, no consistent conclusions relating to format differences have been made in other content areas (e.g., mathematics, science, and language). Therefore, the generalizability of the study was limited to the mathematics examinations only.

However, caution should be exercised when a generalization is applied to other mathematics examinations with different sample size, test length, and structure due to the following two reasons: (1) no research has been done regarding whether the two psychometric models (FIFA and MRCMLM) used in the study are sensitive in detecting dimensionality in varying situations (sample size, test length, and structure); (2) the present study was based on four real data sets, where test length, sample size, and item parameters were not manipulated. Therefore, the conclusion may be better applied to those mathematics examinations that are similar to the four examinations in the present investigation.

CONTRIBUTIONS

The purpose of this investigation was the need to learn if different formats used in the mathematics assessment lead to multidimensionality, and to clarify the nature of the format differences based on the framework of a cognitive theory. Specifically, the study contributed at the level of theory, practice, and methodology. First, the study combined cognitive theory and psychometrical models (MRCMLM and FIFA) to investigate the nature of the differences between the two formats (MC and CR) in assessing students' ability in mathematics. Second, the application of the two multidimensional IRT models had some advantages over other models applied in the previous research studies. Both exploratory (FIFA) and confirmatory (MRCMLM) approaches to the investigation of format problems provided stronger evidence regarding the controversial format issues. Third, the questions of whether the test was unidimensional and whether MC and CR items assessed different mathematical ability for different ability groups of students were answered.

In order to clarify the nature of the format differences, the Bloom's learning taxonomy and the cognitive process models in mathematics problem solving (Mayer, 1985) were applied. In the previous studies, few attempts have been made by the investigators who defined the cognitive skills when comparing MC and CR formats (Hancock, 1996). The cognitive demand analyses helped to explain why several items failed to load with other items on the same factor in the one-factor solution and why several items loaded on the non-significant minor dimensions.

The application of both exploratory and confirmatory factor analyses helped to strengthen the conclusions reached. The FIFA is regarded as the best for exploring dimensionality because it is the combination of both factor analysis and item response theory (Bock et al., 1999). The MRCMLM allows researchers to specify the customized models for their own needs. Using both models is a better approach than the single method applied in the previous studies because both consistent and inconsistent results may provide richer evidence regarding the controversial issues.

The conclusion that MC and CR items measured similar constructs in the present investigation indicated that both formats can be calibrated using unidimensional IRT models. Such finding is important to those teachers and assessment specialists who are in favor of using only certain item format. The hypothesis that high and low ability students differed in handling MC and CR items was supported. Such findings have important implications for assessment specialists when they make decisions as to which item type and how many of them should be included. Such decisions will certainly be very critical for examinations such as those for college entrance or high school graduation.

DIRECTIONS FOR FUTURE RESEARCH

Four types of research studies are necessary in order to investigate the differences between MC and CR items. First, studies should be done in other content areas such as science, social studies, or communication arts. Second, future format comparison studies should be done by manipulating some variables (test length, sample size, and test characteristics). Third, some other models may reveal additional information regarding format differences. Fourth, in order to explore further the nature of the differences between the different formats, it will be useful if the examinees are interviewed about how they deal with the different formats.

According to the previous research studies, the conclusion that MC and CR items differ has been reached only in the domain of writing (Werts et al., 1980; Quellmalz et al., 1982; Ackerman & Smith, 1988). Equivocal results were obtained in science, language arts, and quantitative domains. Because only mathematics examinations were investigated in the present study, similar studies need to be done in the other areas so that generalizations regarding format differences across different domains may be obtained.

The current investigation was based on four real mathematics examinations, which is one factor that may limit the generalizability of the study. Therefore, in future studies, variables such as test length, sample size, or difficulty level can be manipulated so that results can be more generalizable. For example, the results from the present investigation seemed to reveal that whether the test structure is unidimensional for a low ability or high ability group may depend on the difficulty level of the test. Therefore, in future studies, the difficulty variable can be manipulated.

The conditional item pair covariance-based dimensionality test (e.g., DIMTEST) (Stout, 1987) is a nonparametric technique that can be used to explore format differences in the future. This nonparametric technique, which avoids strong parametric modeling assumptions, can be used to explore test structure. Due to the differences that exist between parametric (FIFA and MRCMLM) and nonparametric techniques (DIMTEST), this nonparametric technique may provide additional information.

The cognitive processes that the examinees may engage in when they are doing MC and CR items may be revealed more clearly using "think aloud" procedures. By asking the examinees to report their thought processes as they are working on each item, researchers might be able to obtain more or less direct evidence about the nature of the format differences (Traub, 1993). Analysis of the cognitive processes of examinees may assist in the interpretation of statistical analyses and provide additional information regarding the nature of format differences.

REFERENCES

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement, 12*, 117-128.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67-91.

Ackerman, T. A., Simpson, M. A., & De La Torre, J. (2000, April). *A comparison of the dimensionality of TOEFL response data from different first language groups.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, Louisiana.

Adams, R. J., & Wilson, M. (1996). A random coefficients multinomial logit: A generalized approach to fitting Rasch models. In Engelhard, G. & Wilson, M. (Eds.), *Objective measurement III: Theory into practice* (pp. 143-166). Norwood: NJ:Ablex.

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1-23.

Anastasi, A. (1970). On the formation of psychological traits. *American Psychologist, 25*, 899-910.

Anderson, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50,* 3-16.

Bennett, R. E., Rock, D. A., Braun, H. I., Frye, D., Spohrer, J. C., & Soloway, E. (1990). *Applied Psychological Measurement, 14,* 151-162.

Bennett, R. E., Ward, W. C., Rock, D. A., & LaHart, C. (1990). *Toward a framework for constructed response items* (RR-90-7). Princeton, NJ: Educational Testing Service.

Bennett, R. E., Rock, D.A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28,* 77-92.

Bennett, R. E. & Ward, W. C. (1993). *Construction versus choice in cognitive measurement,* (preface). Hillsdale, NJ: Lawrence Erlbaum.

Bennett, R. E. (1993). On the meanings of constructed-response. In Bennett, R. E. & Ward, W. C. (Eds.), *Construction versus choice in cognitive measurement* (pp. 143-166). Hillsdale, NJ: Lawrence Erlbaum.

Berger, M. P. F., & Knol, D. L. (1990, April). *On the assessment of dimensionality in multidimensional response theory models.* Paper presented at the meeting of the American Educational Research Association, Boston.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats—it does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11,* 385-395.

Birenbaum, M., Tatsuoka, K. K., & Gutvirtz, Y. (1992). Effects of response format on diagnostic assessment of scholastic achievement. *Applied Psychological Measurement, 16,* 353-363.

Bloom, B. S., Englehart, M. B., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives, handbook I: Cognitive domain.* New York: McKay.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. *Psychometrika, 46,* 443-459.

Bock, R. D., Gibbons, R. D., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement, 12,* 261-280.

Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (1999). *TESTFACT 3 Manual.* Chicago, IL: Scientific Software International Inc.

Boodoo, G. (1993). Performance assessments or multiple-choice. *Educational Horizons, 72* (1), 50-56.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29,* 253-271.

British Columbia Ministry of Educational Skills and Training (1998). *British Columbia provincial examination specification document: British Columbia provincial examination (mathematics 12).* Victoria, British Columbia.

Burket, G. R. (1993). *FLUX program.* Monterey, CA: CTB/McGraw-Hill.

Camili, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16,* 129-147.

Carlson, J. E., & Jirele, T. (1992). *Dimensionality of 1990 NAEP mathematics data.* (ERIC Document Reproduction Service No. ED 346 117).

Carlson, J. E. (1993). *Dimensionality of NAEP instruments that incorporate polytomously-scored items.* (ERIC Document Reproduction Service No. ED 360 368).

Carmines, E. G., & Zeller, R. A. (1979*). Reliability and validity assessment.* Beverly Hills CA: Sage.

Carroll, J. B. (1983). The difficulty of a test and its factor composition revisited. In Wainer, H. & Messick, S. (Eds.), *Principals of modern psychological measurement* (pp. 257-282). Hillsdale NJ: Erlbaum.

Champlain, A. D., & Gessaroli, M. E. (1998). Assessing the dimensionality of item response matrices with small sample sizes and short test lengths. *Applied Measurement in Education, 11* (3), 231-253.

Chang, Y., & Davison, M. (1992). *On test information: a comparison of the multidimensional and unidimensional item response theory.* Paper presented at the International Educational Statistics and Measurement Symposium, Taiwan.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In Sternberg, R. S. (ed.), *Advances in the psychology of human intelligence* (Vol. 1, pp. 34-67). Hillsdale, NJ: Lawrence Erlbaum Associates.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40,* 5-32.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd.). Hillsdale, N. J.: Lawrence Erlbaum Associates.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Belmont, CA: Wadsworth Publication Company.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data with the EM algorithm. *Journal of the Royual Statistical Society, Series B, 34,* 1-34.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: a pragmatic approach* (Research report, 47). Princeton, NJ: Educational Testing Service.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement, 35* (2), 137-154.

Feltovich, P. J. (1981). *Knowledge based components of expertise in medical diagnosis* (Technical Report No. PDS-2). Pittsburg, PA: University of Pittsburgh Learning Research and Development Center.

Ferguson, G. A. (1956). On transfer and the abilities of man. *Canadian Journal of Psychology, 10,* 121-131.

Fiske, E. B. (1990). But is the child learning? Schools trying new tests. *The New York Times,* A1-A6.

Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement, 13,* 373-389.

Frary, R. B. (1995). Multiple-choice versus free-response: a simulation study. *Journal of Educational Measurement, 22,* 21-31.

Gagne, E. D. (1985). *The cognitive psychology of school learning* (Section II-III). Boston: Little Brown Company.

Geeslin, W. E., & Shavelson, R. J. (1975). An exploratory analysis of the representation of a mathematical structure in students' cognitive structures. *American Educational Research Journal 12,* 21-39.

Gorsuch, R. L. (1974). *Factor analysis.* Philadephia, PA: Saunders.

Guilford, J. P. (1971). *The nature of human intelligence.* London: McGraw-Hill.

Hambleton, R. K., & Murray, L. N. (1983). Some goodness of fit investigations for item response models. In Hambleton, R. K. (Ed.), *Applications of item response theory.* Vancouver, BC: Educational Research Institute of British Columbia.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory, principles and applications.* Boston: Kluwer · Nijhoff Publishing.

Hancock, G. R. (1996). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *Journal of Experimental Education, 62* (2), 143-157.

Harke, D. J., Herron J. D., & Lefler, R. W. (1972). Comparison of a randomized multiple-choice format with a written one-hour physics problem test. *Science Education, 56,* 563-565.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioural Research, 19,* 49-78.

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9* (2), 139-164.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In E. W. Gregg, E. W., & Sternberg, E. R. (Eds.), *Cognitive processes in writing* (pp. 3-30). Hillsdale NJ: Erlbaum.

Hendrickson, A. E., & White, P. O. (1964). Promax: a quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology, 17,* 65-70.

Hinsley, D., Hayes, J. R., & Simon, H. A. (1977). From words in equations. In Carpenter, P. & Just, M. (Eds.), *Cognitive processes in comprehension.* Hillsdale, NJ: Erlbaum.

Hogan, T. P. (1981). *Relationship between free-response and choice-type tests of achievement: A review of the literature.* (ERIC Document Reproduction Service No. ED 224 811).

Hulin, C. L., Drasgow, F., & Pearsons, C. K. (1983). *Item response theory: application to psychological measurement.* Homewood, IL: Dow Jones-Irwin.

Hurley, J. R., and Cattell, R. B. (1962). The Procrustes program: producing a direct rotation to test a hypothesized factor structure. *Behavioral Science 7,* 258-62.

Janssen, R., & De Boeck, P. (1996). The contribution of a response-production component to a free-response synonym task. *Journal of Educational Measurement, 33* (4), 417-432.

Joreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance metrices. In Atkinson, R. C., Krantz, D. H., Luce, R. D., & Suppes, P. (Eds.), *Contemporary Developments in Mathematical Psychology* (Vol. 2, pp. 1-56). San Francisco: W. H. Freeman.

Joreskog, K. G., & Sorbom, D. (1978). *LISREL 7 user's reference guide.* Mooresville, IN: Scientific Software, Inc.

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika, 35,* 401-415.

Kelley, T. L. (1938). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30,* 17-24.

Loftus, E. F., & Suppes, P. (1972). Structural variables that determine problem-solving difficulty in computer-assisted instruction. *Journal of Educational Psychology, 63,* 531-542.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading Mass: Addison-Wesley.

Lord, F. M., (1971). *Testing if two measuring procedures measure the same psychological dimension* (Research Bulletin RB-71-36). Princeton, NJ: Educational Testing Service.

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

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed-response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement, 31,* 234-250.

Martinez, M. E. (1991). Comparison of multiple-choice and constructed figural response items. *Journal of Educational Measurement, 28,* 131-143.

Mayer, R. E. (1980). Schemas for algebra story problems. *Series in Learning and Cognition* (Report No, 80-3). Santa Barbara, CA: University of California, Department of Psychology,

Mayer, R. E. (1982). Memory for algebra story problems. *Journal of Educational Psychology, 74,* 199-216.

Mayer, R. E. (1985). Mathematical ability. In Sternberg, R. J. (Ed.), *Human abilities: an information processing approach* (pp. 127-150). San Francisco: Freeman.

McDonald, R. P. (1967). *Non-linear factor analysis* (Psychometric Monograph No.15). Iowa City: Psychometric Society.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6,* 379-396.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11,* 3-31.

Mislevy, R. J. (1993). A framework for studying differences between multiple-choice and free-response test items. In Bennett, R. E., & Ward, W. C. (Eds.), *Construction versus choice in cognitive measurement* (pp. 75-106). Hillsdale, NJ: Lawrence Erlbaum.

Muraki, E., & Engelhard, G. J. (1985). Full-information item factor analysis: applications of EAP scores. *Applied Psychological Measurement, 9* (4), 417-430.

Muthen, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43,* 551-560.

Muthen, B. (1984). A general structural equation model with dichotomous, ordered categories, and continuous latent variable indicators. *Psychometrika, 49,* 115-132.

Nickerson, R. S. (1989). New directions in educational assessment. *Educational Researcher, 18*, 3-7.

Ormell, C. P. (1974). Bloom's taxonomy and the objectives of education. *Educational Research, 17* (1), 3-18.

O'Neil, H. F., & Brown, R. S. (1998). Differential effects of question formats in math assessment on metacognition and affect. *Applied Measurement in Education, 7 (*4), 331-351.

Quellmalz, E. S., Capell, F., & Chou, C. P. (1982). Effects of discourse and response mode on the measurement of writing competence. *Journal of educational measurement, 19,* 241-258.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics, 4,* 207-230.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21* (1), 25-36.

Rogers, H. J. (1984). *Fit statistics for latent trait models.* Unpublished master's thesis. University of New England, Armidale, Australia.

Samejima, F. (1972). A general model for free-response data. *Psychometric Monograph, 18.*

Silver, E. A. (1981). Recall of mathematical problem information: solving related problems. *Journal for Research in Mathematics Education, 12,* 54-64.

Smith, R. M., & Miao, C. Y. (1994). Assessing unidimensionality for Rasch measurement. In Wilson, M.(Ed.), *Objective measurement: theory into practice* (pp. 316-327). Norwood, NJ: Ablex Publishing Corp.

Snow, R. E. (1993). Construct validity and constructed-response tests. In Bennett, R. E. & Ward, W. C. (Eds.), *Construction versus choice in cognitive measurement*_(pp. 45-60). Hillsdale, NJ: Lawrence Erlbaum.

Snow, R. E., & Lohman, D. F. (1993). Implications of cognitive psychology for educational measurement. In Robert, L. L. (Ed.), *Educational Measurement* (pp. 263-331). NY: Macmillan Publishing Company.

Stout, W. F. (1987). A new item response theory modelling approach with applications to unidimensional assessment and ability estimation. *Psychometrika, 52,* 589-617.

Third International Mathematics and Science Survey Technical Report (Vol. 1) (1996). Boston College, Chestnut Hill, Massachusetts.

Thissen, D., Wainer, H., & Wang, X. B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement, 31,* 113-123.

Traub, R. E., & Fisher, C. W. (1977). On the equivalence of constructed response and multiple choice tests. *Applied psychological measurement, 1* (3), 355-369.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In Hambleton, R. K. (Ed.), *Applications of item response theory* (pp. 57-70). Vancouver, BC: Educational Research Institute of British Columbia.

Traub, R. E., & MacRury, K. (1990). *Multiple-choice vs. free-response in the testing of scholastic achievement.* Weinheim, Germany: Beltz Verlag.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In Bennett, R. E. & Ward, W. C. (Eds.), *Construction versus choice in cognitive measurement* (pp. 29-44). Hillsdale, NJ: Lawrence Erlbaum.

Travers, R. M. W. (1980). *Taxonomies of educational objectives and theories of classification. Educational Evaluation and Policy Analysis, 2* (2), 5-23.

Van den Bergh, H. (1990). On the construct validity of multiple-choice items for reading comprehension. *Applied Psychological Measurement, 14,* 1-12.

Volodin, N. A., & Adams, R. J. (1995). *Identifying and estimating a d-dimensional item response model.* Paper presented at the International Objective Measurement Workshop, University of California, Berkeley, California.

Walker, C., & Beretvas, S. N. (2000, April). *Using multidimensional versus unidimensional ability estimates to determine student proficiency in mathematics.* Paper presented at the 2000 Annual Meeting of the American Educational Research Association, New Orleans, Louisiana.

Wallenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47*, 123-140.

Wang, W. C. (1994). *Implementation and application of the multidimensional random coefficients multinomial logit model.* Unpublished doctoral dissertation, University of California, Berkeley.

Wang, W. C., & Wilson, M. (1996). Comparing multiple-choice items and performance-based items using item response modelling. In Engelhard, G. & Wilson, M. (Eds.), *Objective measurement III: Theory into practice* (pp. 167-194). Norwood, NJ: Ablex.

Ward, J. H., (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58,* 236-244.

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement, 6,* 1-11.

Werts, C. E., Breland, H. M., Grandy, J., & Rock, D. R. (1980). Using longitudinal data to estimate reliability in the presence of correlated measurement errors. *Educational and Psychological Measurement, 40,* 19-29.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wu, M. L. (1997). *The development and application of a fit test for use with marginal maximum likelihood estimation and generalised item response models.* Unpublished master's thesis. University of Melbourne, Melbourne, Australia.

Wu, M. L., Adams, R. J., & Wilson, M. (1997). *ConQuest, generalized item response modelling software* (draft release 2). Australian Council for Educational Research.

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement, 30* (3), 187-213.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement, 24* (4), 293-308.

APPENDIX A: DATA SET ONE

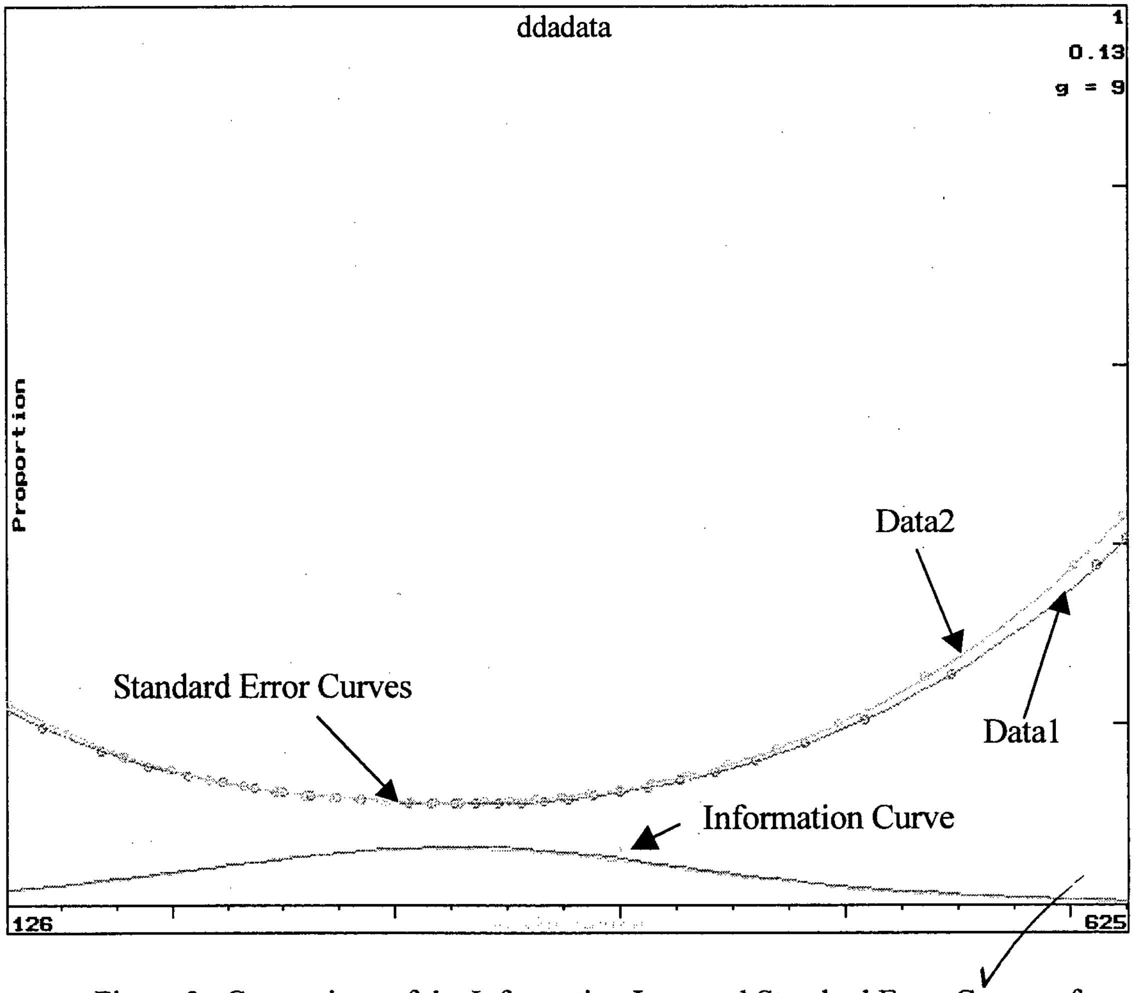


Figure 3. Comparison of the Information Loss and Standard Error Curves of Two Data (Study One). 1—Original Data, 2—Dichotomized Data

```
2                                           |17
                                            |
                                            |
                      XXXXXXXXXXXXXXXXXXXXXX|
                                            |
                                      XXXXXX|
                                            |
                       XXXXXXXXXXXXXXXXXXXXX|
1                              XXXXXXXXXXXXX|12
                XXXXXXXXXXXXXXXXXXXXXXXXXXXX|16 21 27 29
                                  XXXXXXXXXX|
                     XXXXXXXXXXXXXXXXXXXXXXX|25
                     XXXXXXXXXXXXXXXXXXXXXXX|6 28
                      XXXXXXXXXXXXXXXXXXXXXX|15
                                   XXXXXXXXX|8 11 24
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|18
0                          XXXXXXXXXXXXXXXXX|2 3 13 20
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|10 19 26
                           XXXXXXXXXXXXXXXXX|9
                                     XXXXXXX|
                          XXXXXXXXXXXXXXXXXX|
                    XXXXXXXXXXXXXXXXXXXXXXXX|
                                     XXXXXXX|
-1                                          |23
                        XXXXXXXXXXXXXXXXXXXX|
                                            |22
                                            |1 5
                               XXXXXXXXXXXXX|7
                                            |
                                       XXXXX|
                                          XX|4 14
-2                                          |
                                    XXXXXXXX|
                                   XXXXXXXXX|
                                      XXXXXX|
                                            |
-3
```

Items plotted at their difficulty estimates.

Figure 4. Map of latent distribution and response model parameter estimates for Unidimensional Model (Data Set One). Bold items are constructed-response items, non-bold items are multiple-choice items (29 items from TIMSS, Gr.3 and Gr.4 Mathematics Examination).

APPENDIX B: DATA SET TWO

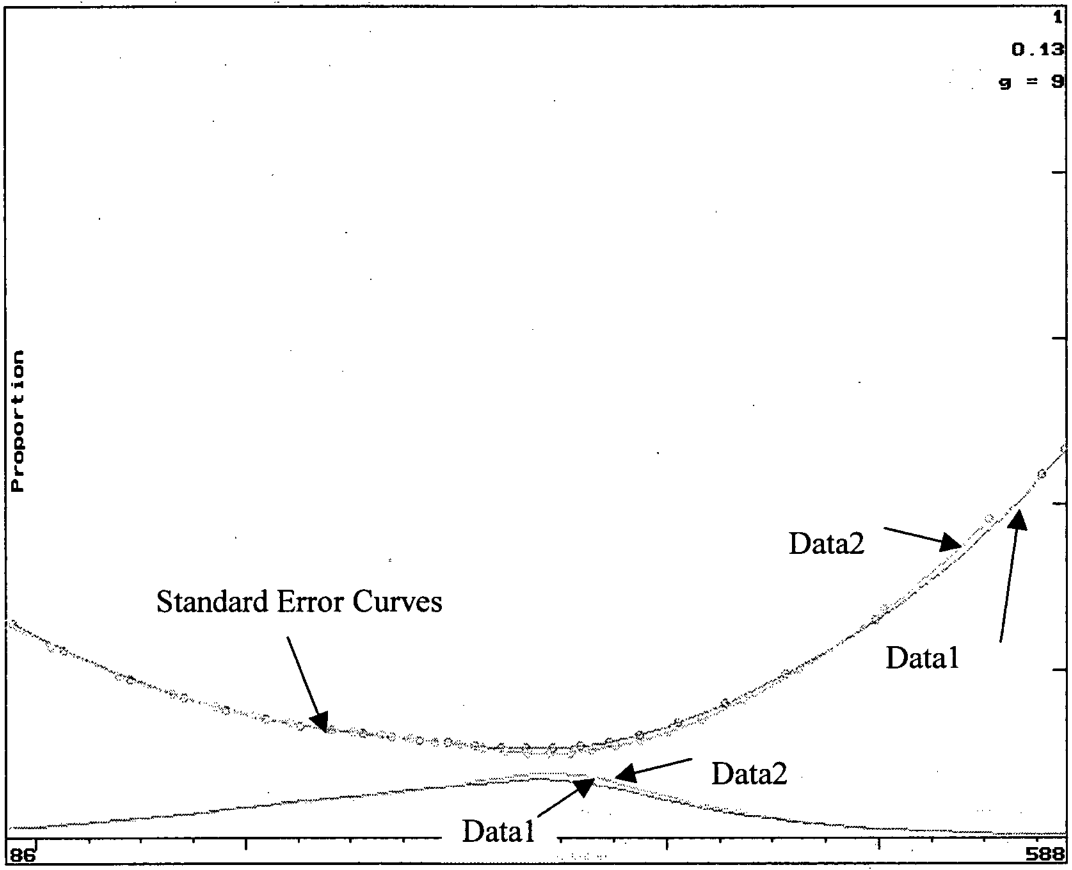
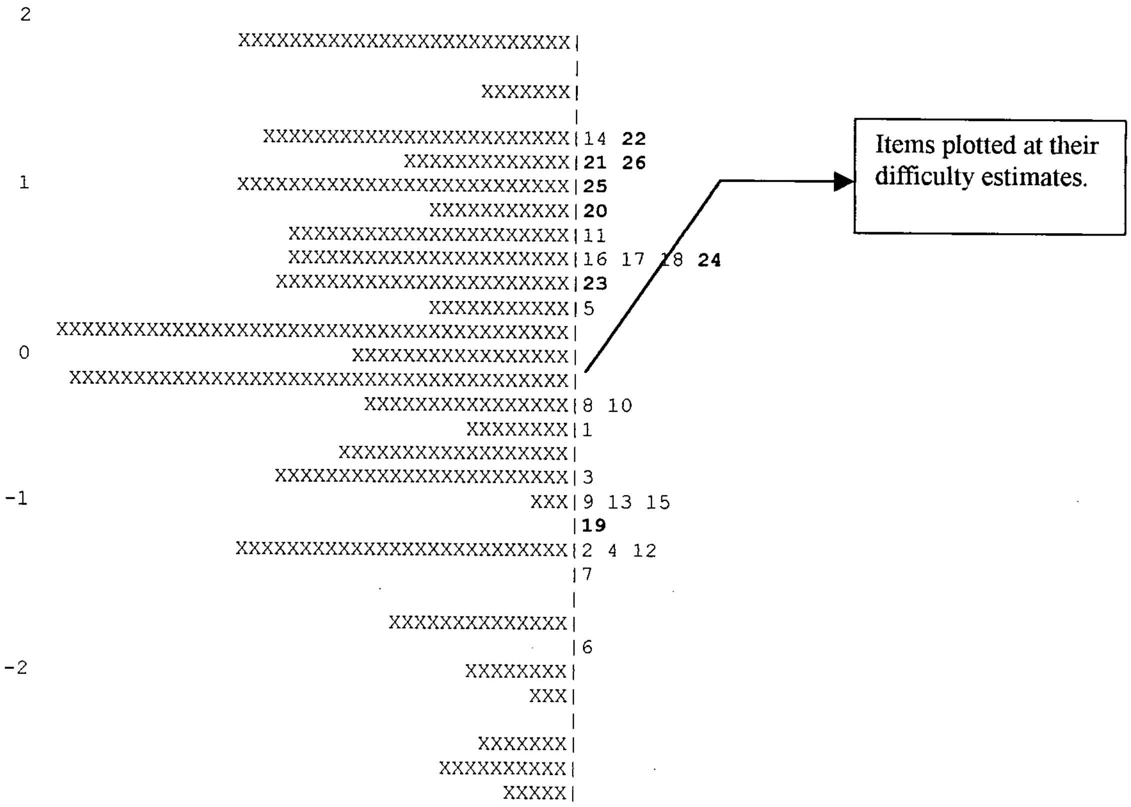


Figure 7. Comparison of the Information Loss and Standard Error Curves of Two Data (Data Set Two). 1—Original Data, 2—Dichotomized Data



Items plotted at their difficulty estimates.

Figure 8. Map of Latent Distribution and Response Model Parameter Estimates for One-Dimension Format Model (Data Set Two). Bold items are constructed-response items, non-bold items are multiple-choice items (26 items from TIMSS, Gr.7 and Gr.8 Mathematics Examination).

APPENDIX C: DATA SET THREE

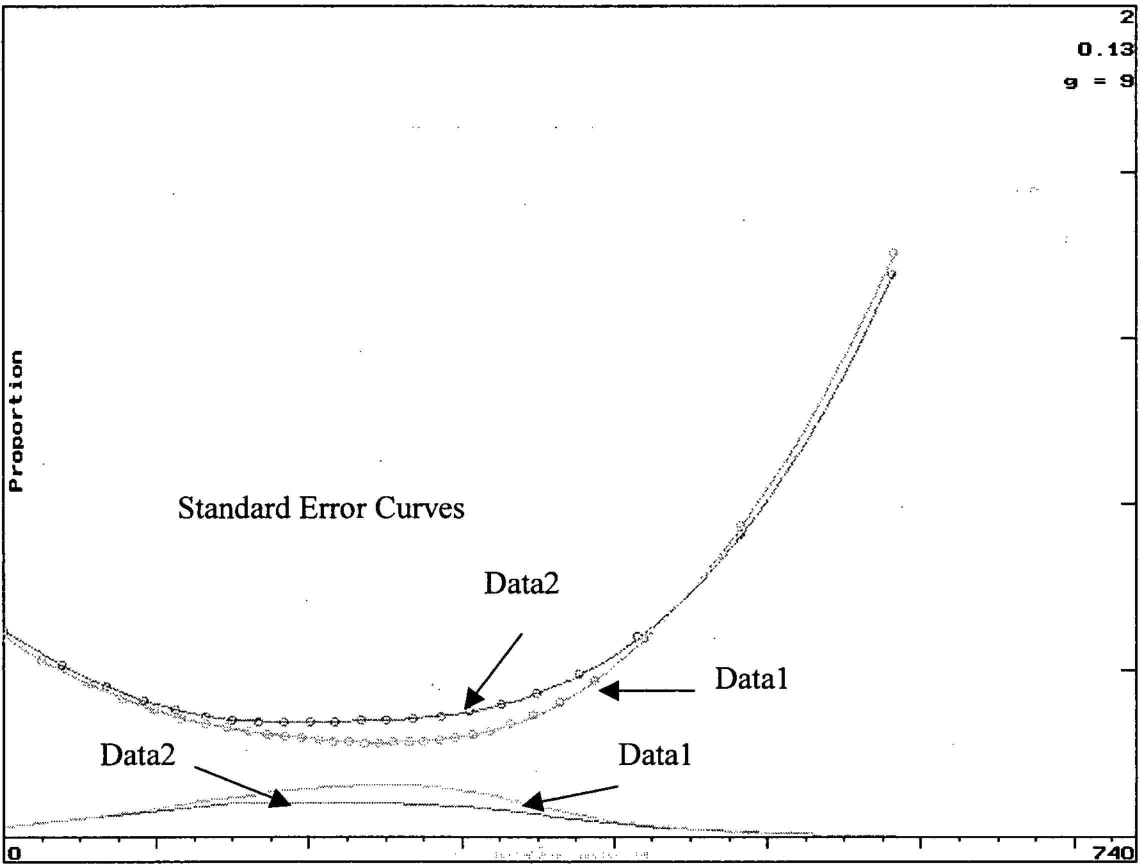


Figure 11. Comparison of the Information Loss and Standard Error of Two Data (Study Three). 1—Original Data, 2--Dichotomized

```
  2
                         XXXXXXXXXXXXXXXXXXXXXXX|
                                                |
                                       XXXXXXXXX|
                                                |
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXX|
                                XXXXXXXXXXXXXXXX|15
  1             XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|
                                       XXXXXXXXX|
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXX|
                       XXXXXXXXXXXXXXXXXXXXXXXXX|
                   XXXXXXXXXXXXXXXXXXXXXXXXXXXXX|18  23
                                    XXXXXXXXXXXX|21
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|
  0                        XXXXXXXXXXXXXXXXXXXXX|6  9
               XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|3
                                  XXXXXXXXXXXXXX|19
                                      XXXXXXXXXX|
                   XXXXXXXXXXXXXXXXXXXXXXXXXXXXX|
                        XXXXXXXXXXXXXXXXXXXXXXXX|
 -1                                      XXXXXXX|17
                                                |14
                           XXXXXXXXXXXXXXXXXXXXX|12
                                                |
                                                |
                                XXXXXXXXXXXXXXXX|2
                                                |5
 -2                                         XXXX|11  22
                                            XXXX|20
                                                |10
                                           XXXXX|
                                        XXXXXXXX|
                                          XXXXXX|
                                                |
 -3                                             |4  13  16
                                                |1
                                                |7
                                                |
                                                |8
                                                |
```

Items plotted at their difficulty estimates.

Figure 12. Map of Latent Distribution and Response Model Parameter Estimates for Unidimension Model (Data Set Three). Bold items are constructed-response items, non-bold items are multiple-choice items (23 items from British Columbia Grade12 Mathematics Examination, April 1998).

APPENDIX D: DATA SET FOUR

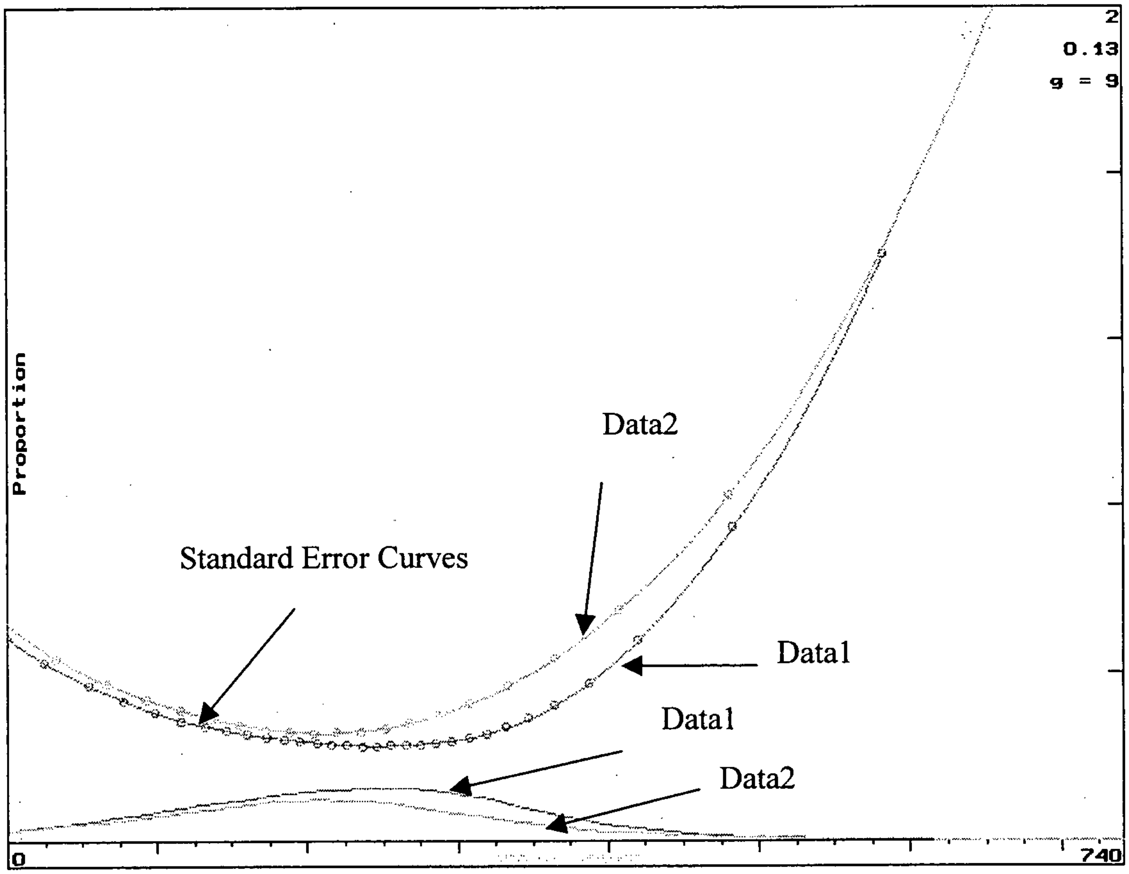


Figure 15. Comparison of the Information Loss and Standard Error Curves of Two Data (Data Set Four). 1—Original Data, 2—Dichotomized Data.

```
  2
                      XXXXXXXXXXXXXXXXXXXXXXX|
                                             |
                                    XXXXXXXXX|
                                             |
                 XXXXXXXXXXXXXXXXXXXXXXXXXXXX|
                             XXXXXXXXXXXXXXXX|15
  1          XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|
                                    XXXXXXXXX|
                 XXXXXXXXXXXXXXXXXXXXXXXXXXXX|
                    XXXXXXXXXXXXXXXXXXXXXXXXX|
                XXXXXXXXXXXXXXXXXXXXXXXXXXXXX|18 23
                                 XXXXXXXXXXXX|21
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|
  0                     XXXXXXXXXXXXXXXXXXXXX|6 9
            XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|3
                               XXXXXXXXXXXXXX|19
                                   XXXXXXXXXX|
                XXXXXXXXXXXXXXXXXXXXXXXXXXXXX|
                     XXXXXXXXXXXXXXXXXXXXXXXX|
 -1                                   XXXXXXX|17
                                             |14
                        XXXXXXXXXXXXXXXXXXXXX|12
                                             |
                                             |
                             XXXXXXXXXXXXXXXX|2
                                             |5
 -2                                      XXXX|11 22
                                         XXXX|20
                                             |10
                                        XXXXX|
                                     XXXXXXXX|
                                       XXXXXX|
                                             |
 -3                                          |4 13 16
                                             |1
                                             |7
                                             |
                                             |8
                                             |
```

Items plotted at their difficulty estimates.

Figure 16. Map of Latent Distribution and Response Model Parameter Estimates for Unidimension Model (Data Set Four). Bold items are constructed-response items, non-bold items are multiple-choice items (23 items from British Columbia Grade12 Mathematics Examination, August 1998).