This is to certify that the
dissertation entitled

EFFECTS OF TEST LINKING METHODS ON PROFICIENCY
CLASSIFICATION: UIRT VERSUS MIRT LINKING

presented by

YOUNG YEE KIM

has been accepted towards fulfillment
of the requirements for the

Doctor of Philosophy degree in Measurement and Quantitative Methods Program
Educational Policy Program

Mark D. Reckase
Major Professor's Signature

Jan. 7, 2008

MSU is an affirmative-action, equal-opportunity employer

EFFECTS OF TEST LINKING METHODS ON PROFICIENCY CLASSIFICATION: UIRT VERSUS MIRT LINKING

By

Young Yee Kim

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Measurement and Quantitative Methods
Educational Policy

2008

ABSTRACT

EFFECTS OF LINKING METHODS ON PROFICIENCY CLASSIFICATION: UIRT VERSUS MIRT LINKING

By

Young Yee Kim

The major purpose of this study was to show educational policy implications of psychometric decisions on educational measurement by exploring the effects of selecting different linking approaches – UIRT vs. MIRT linking – on proficiency rate changes. The result shows that different linking approaches and different choices of proficiency classification models produce different conclusions on the educational progress inferred from increased or decreased proficiency rates across years.

In this study, a fixed common item parameter (FCIP) linking method was selected to link two years' test data in both the UIRT and MIRT linking approaches. Five random sample data (RSD) sets of 10,000 samples were selected for two years and each RSD set of one year (2005) was linked to its matched RSD set of 2006. For an UIRT approach, the PARSCALE program was employed. The two years' test data were calibrated separately by running PARSCALE and 2005 results were recalibrated by fixing the common items with the item parameters calibrated in 2006 data. For a MIRT approach, each RSD set was calibrated by running the BMIRT Program. For FCIP MIRT linking, the BMIRTanchor program was employed.

This study found that different linking approaches and different decision approaches to proficiency classification produced different results. Overall, the UIRT approach was favorable to the 2006 students. While there was little change in proficiency rate between 2005 and 2006 using the UIRT approach for both of the classification

criteria, the MIRT compensatory and conjunctive approaches resulted in a decreased proficiency rate in 2006 compared to 2005 for the 20th percentile classification criterion and no statistically significant difference for the 50th percentile classification criterion. This result strongly suggests the importance of selecting a linking method and a proficiency classification approach when evaluating educational progress by the change of proficiency rate.

ACKNOWLEDGEMENTS

There are so many people who I owe a lot throughout my doctoral study in the College of Education. While I cannot name all of them individually, it is my great honor and pleasure to express my deep appreciation and gratitude to some of them.

First, I'd like to express my sincere and deep appreciation to my advisor for the Measurement and Quantitative Methods Program and the dissertation chair, Dr. Mark Reckase. He has been more than an advisor and dissertation chair especially during the last three years when I have become a true lover of psychometrics and multidimensional item response theory. He has been a best friend and mentor by providing me with the best emotional support as well as the best academic guidance and support. I have admired not only his profound knowledge and skills in psychometrics and willingness to share them with his students but also his insights and deep understanding of educational policy issues related to measurement and assessment. His willingness to help and his appreciation of my dissertation research has been the most important source of my strengths to complete my dissertation study. He has been there whenever I needed his help and guidance. I have been truly blessed with having the wonderful opportunities to work with and learn from him.

I'd also like to express my sincere and deep appreciation to my academic advisor for the Educational Policy Program, Dr. Gary Sykes. Meeting with him was a turning point in my new academic journey in Educational Policy. Even before serving as my academic advisor, he provided me with his best guidance and advice whenever I needed them. His recognition of my academic capability helped me to regain self-confidence in my academic ability when I was struggling to adjust to a new academic environment. Gary also served on my dissertation committee. I appreciate his willingness to serve on the committee and his emotional support and his insightful comments on the educational policy implications of my study.

Two other committee members, Dr. Joan Ferrini-Mundy and Sharif Shakrani, shared their time with me in the middle of their busy schedules. I was truly fortunate to have them on my dissertation committee. Joan provided valuable comments on the relevance of my study to mathematics education assessment policy and encouraged me to elaborate my discussion of mathematical constructs, which I believe extended the educational implications and importance of my study. Sharif helped me to clarify my arguments and to extend the educational policy relevance and implications of my study.

Other faculty members at Michigan State University also helped me during my doctoral study. Dr. Jack Schwille has been my mentor, someone from whom I could ask advice whenever I needed help and support. Dr. Jon Star (now in Harvard University) has been a good friend and mentor, too. I am especially thankful to his consistent support and help while conducting my apprenticeship research project. I also like to express my appreciation to Dr. Raven McCrory who was my supervisor for my graduate research assistantship for the last two years. Through working with her, I gained useful knowledge, skills, and experiences in educational research. I am also thankful for the support and help I have received from Dr. Peter Youngs, who helped me in many occasions, reading the drafts of my presentation papers and giving advices in relation to my job search. Joan Smith was my supervisor for my graduate assistantship for a year and then became a good friend. I thank Joan for her willingness to give me a company when I needed. I would also like to express my thanks to professors in the MQM program and the Educational Policy program for sharing their knowledge and expertise in their own areas and other professors in the College of Education, especially, Dr. Jack Smith for his kindness to share his research project room with me when I needed a space to work on campus.

I would like to express my special thanks to Dr. Michael Sedlak and Dr. Ken Frank for encouraging me to apply to a doctoral dissertation completion fellowship as well as to the Graduate School at Michigan State University for providing the fellowship for me.

All of my Korean friends at MSU who shared our agony and pleasure of pursuing doctoral study and conducting research, especially Ji-Won Son, Rae-Young Kim, Ga-Young Ahn, and Yonghee Suh, deserve my thanks. I also thank my writing group members—Michael Sherry, Sungworn Ngudgratoke, and Dipendra Subedi at MSU— for sharing their time and ideas and giving me valuable comments, especially during the conception stage of my dissertation. My thanks also go to Kosze Lee, Joo-Young Park, Violeta Yurita, and Marcy Wood who served on the Item Review Committee for my dissertation study. I appreciate Young Soon Shin, my lifetime friend, for her consistent encouragement and kind friendship.

I am thankful to Educational Quality and Accountability Office in Ontario, Canada and Dr. Michael Kozlow for allowing me to use the data for my study and also to the staff at EQAO for their assistance and help in the process of getting all the data I needed.

I also appreciate Dr. Lihua Yao at CTB/McGraw-Hill for providing BMIRT programs and technical advices and help in relation to running the programs.

I am happy to have an opportunity to express my appreciation to wonderful colleagues at my workplace, NAEP- Educational Statistics Service Institutes at American Research Institutes; especially Cadelle Hemphill, Enis Dogan, Natalia Pane, and Alan Vanneman for their supports in working on my dissertation while beginning a new job. I especially appreciate Alan for sparing his time to read the draft of my dissertation and providing many valuable comments. In addition to editorial comments, he made me reread my writing and clarify my expressions by giving me feedback from the perspective of a "lay" person with little psychometric knowledge. He also taught me how to make text smoother and clearer with small revisions.

Finally but most importantly, I'd like to express my deepest appreciation to my family; First, to my trustful husband, Sangki Han, and my lovely daughter, Min-Hye Han, for their love and trust in me and consistent support during the whole process of my pursuing dual major doctoral degrees. They have provided me with wonderful support and help. I also appreciate their considerate understanding of my failure to spend together as much time as they would like to while I was working on my dissertation. I owe them tremendously; Second, to my parents in law and brothers and sisters in law for their support and trust in me; Last, to my late parents for their unconditional love and complete trust in me, and my sisters and brothers for their consistent and steadfast support and love for me.

TABLE OF CONTENT

CHAPTER 4 RESULTS AND INTERPRETATION

CHAPTER 5 SUMMARY, DISCUSSION, AND CONCLUSION

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Improving students' academic achievement is one of the major goals of the current educational policy across the world. Large-scale assessments have been used as "levers" to implement standards-based educational reform to improve students' academic achievement in the United States. Teaching and thus learning is expected to be aligned with rigorous, high quality academic standards. Test scores are believed to indicate the amount of students' learning when the tests are aligned with the standards. Test scores are used as a tool to hold schools and educational system accountable for educational outcomes (Darling-Hammond, 2002; Linn, 2003). Accountability policy based on assessment culminated in the adequate yearly progress (AYP) requirement of No Child Left Behind (NCLB) Act (2001) in the case of the United States.

The professed goal of NCLB is to make all students achieve at or above "proficiency" performance levels for reading and mathematics set by each state by the 2013-2014 school year. This goal of 100 percent proficiency by 2014 has been criticized as being "unrealistic" though "commendable" (Linn, 2005). While there are many components to NCLB, such as the regulation of teacher quality and professional development for capacity building, the integral part is accountability and assessment policy. That is, educational systems and schools are required to be accountable for educational achievement (i.e., the goal of NCLB) and thus the change (i.e., increase of proficiency rate) has become one important measure of accountability.

Educational policy centered on accountability and assessment to improve educational achievement is not unique to the United States. For example, the Council of Ministries of Education, Canada (CMEC), which was founded in 1967 by ministers of education in provinces and territories in Canada, administered a program of pan-Canadian assessments of student achievement in mathematics, reading and writing, and science – the School Achievement Indicators Program (SAIP) – between 1993 and 2004. CMEC recently replaced the program with the Pan-Canadian Assessment Program (PCAP), which continues to assess performance in the same three core subjects as SAIP[1]. Also, province-wide achievement testing of elementary and high school students has been the norm throughout Canada (Volante, 2004). In Canada, however, educational policy and setting educational goals are the responsibility of each province and territory. Educational goals, therefore, vary across provinces. Some provinces have more "realistic" educational goals than those specified in the NCLB compared to the United States. For example, Ontario province has set its educational goal as 75% above the proficiency criterion.

In response to educational policy efforts to improving educational achievement through the lever of large-scale assessment tests, some concerns about, and criticisms of, using test scores from large-scale assessment as a measure of the performance of an educational system have been raised. Concerns about the role of large-scale assessment in improving students' achievement can be classified as three categories.

First, there is a question of whether large scale assessments can really measure students' learning in areas or dimensions which subject-specific educators consider important. The question of whether such assessment policy really contributes to

[1] Source: http://www.cmec.ca/index.en.html

improving “genuine or authentic” learning has been raised by some members of the education community (e.g., Stake, 1995; Volante, 2006). Some scholars are doubtful that educational policy centering on assessment will bring increased genuine learning. One of the major concerns is that the focus of instruction would be given to “teaching to the test” (Volante, 2004). The second concern is whether large-scale assessments provide diagnostic information to educators and teachers that they can utilize to improve students’ learning. Third, there is a question of whether test scores are meaningfully comparable across years. This question is directly related to linking methods or procedures which are designed to place test scores from different test forms on a common (i.e., comparable) scale.

1.2 Concerns Regarding Large-Scale Mathematics Assessments

In the case of the mathematics education community, their concerns in relation to the assessment policy can be summarized in two categories.

First, some mathematics educators say that large-scale standardized achievement tests do not measure important mathematical knowledge and skills such as conceptual understanding or mathematical reasoning. They are particularly critical of tests that consist mainly of multiple-choice (MC) questions. In response, some states are introducing mixed format tests, i.e., tests consisting of both dichotomous items (multiple-choice items) and polytomous items (open-response or constructed-response questions). A recent survey reports that 63% of the state assessments use both multiple-choice items and constructed-response items[2] (Lane, 2005).

[2] Depending on the testing program, items which require students to write answers are called “open-response” items, (i.e., the testing program in Ontario Province, Canada) or constructed-response items (i.e.,

It is believed that constructed-response items measure different knowledge and skills than those measured by multiple-choice items (Traub, 1993). Also some research suggests that constructed-response items provide more information than other type of items (Donoghue, 1993). Common sense suggests that open-response items in mathematics tests provide more specific measurement of students' mathematical knowledge and skills, but whether mixed format tests consisting of both dichotomous items[3] and polytomous items[4] can provide better information about student learning needs more empirical research. If empirical research can show the relative advantage of different format items in terms of measuring mathematical knowledge and skills more accurately, arguments for including constructed-response items would be more convincing.

A second criticism of large-scale assessment tests from mathematics educators is that such tests do not provide diagnostic information useful for instructional improvement and thus students' learning. For example, when students do not achieve the proficiency level, what do test scores say about instructional strategies to help them with their weakness? In relation to "diagnostic" information, NCLB requires states to report diagnostic scores for each content sub-domain or strand (Horton & Hanes, 2005). In the case of mathematics, many states usually report either mean scale scores or percent correct scores by content strand such as algebra, geometry, measurement, etc. This approach has been criticized because of the potential low reliability of subscale scores

in the National Assessment of Educational Progress). These terms are interchangeably used in this dissertation.

[3] They refer to multiple choice question items which are scored dichotomously either "correct" (=1) or "incorrect (=0).

[4] They refer to constructed response items which have more than two (1 or 0) scoring categories.

due to small numbers of items per content area (Haberman, 2005; Monaghan, 2006; Sinharay, Haberman, & Puhan, 2007).

There are two main problems with the practice of reporting for sub-scores by number correct scores. One is the assumption behind this reporting practice that the score for each content strand reflects achievement in the content strand area. Because typical educational achievement test items require students to be "proficient" in more than one ability dimension in order to provide a correct response, this assumption is difficult to support. For example, algebra items in the early grades (grade 6, for example) typically require computational ability as well as algebraic knowledge and skills. Proficiency in the algebra content strand measured by the sub-score from "algebra" items according to a test specification might reflect both computational ability and algebraic ability. Whether or not reporting sub-scale score by content strand is a psychometrically valid procedure has not been fully investigated. Some research efforts at improving subscore reporting by addressing the problem of reporting number correct scores per content strand (in the case of mathematics) or content area have been made by utilizing a Bayesian augmented approach to address the reliability issues due to small number of items per sub-score reporting areas or subscales (Thissen & Edwards, 2005; Edwards & Vevea, 2006; Yao & Boughton, 2007). These researchers have tried to provide better estimates of subscale scores than number-correct scores, either utilizing MCMC estimation procedures (Edwards & Vevea, 2006; Yao & Boughton, 2007) or a multivariate generalization of Kelly's classic regressed estimate of the true score (Thissen & Edwards, 2005). However, estimates of subscores by these approaches are still based on classification of items by test specification while research shows that test items are not always classified as the same content category as that specified by test specification. For example, Herman, Webb, and Zuniga (2005) reported that the kappa coefficients for assignment of items to content categories were .71 and .74 for faculty and teacher raters, respectively. This suggests that items can be classified into different content categories than those specified by test specification across different raters. Then, treating the classification of items into content strands or categories by test specification as fixed truth becomes problematic.

Second, when mathematical ability is reported by content strand, there is no way of measuring mathematical ability in terms of mathematical process dimensions such as problem solving, mathematical reasoning, or communication etc., which has been defined as one of the major goals of mathematics education (NCTM, 2000).[5]

1.3 Concerns Regarding Reporting Educational Achievement through Linked or Equated Score Scales

Even if it is assumed that tests can measure student achievement accurately, the question of whether improvement of student achievement can be validly inferred based on test scores remains to be addressed. To interpret the increase or decrease in the percentage of proficient students as improvement or regression in student achievement, it is necessary to ensure the comparability of test scores across years. For example, if the difficulty of tests varies across years, it will not be possible to compare the test scores from different test forms even though tests across years are based on the same test specifications.

[5] *Principles and Standards for School Mathematics* (NCTM, 2000) presents five process standards for school mathematics: Problem Solving, Reasoning and Proof, Communication, Connections, and Representation.

If the same test is administered each year, test scores and thus percentage of proficient students across years can be compared directly to measure educational achievement. However, it is often not possible to administer the same test each year because test items are released for the purpose of helping guide the educational system and also to meet public interest and concern about the educational testing system. It is claimed that tests consisting of different items are still measuring the same construct, so that the test results can be compared. If two tests are very similar in contents and constructs being measured, but differ a little in difficulty, test scores can be put on the same metric through a statistical procedure called test equating. When there are multiple test forms for security reasons, for example in high stakes tests such as the SAT or ACT[6], test scores from different forms can also be put on the same metric through test equating so that test scores across forms can be compared directly.

The issue of validity of inferences on educational achievement based on test scores has provided increased attention to the equating procedures. Kolen and Brennan (2004) indicate that awareness of the importance of equating has increased among measurement professionals and test users in response to arguments advanced by testing critics in the context of the accountability movement in education.

Equating is defined as "a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably" (Kolen & Brennan, 2004). For test forms to be equated, they should be constructed based on the same test specifications with the same construct(s) and contents as well as similar difficulty. If

[6] SAT and ACT are standardized tests for college admissions in the United States. For more information about SAT and ACT, refer to the websites of the testing programs, www.collegeboard.com and www.actstudent.org/.

equating is successfully done, test results can be validly compared, so that it is possible to talk about score gains and thus improvement in educational achievement.

In the case of educational achievement tests, however, tests might need to be changed across years to reflect change in curriculum frameworks or instructional emphasis and practices. When tests across years change in format and content emphasis in addition to difficulty, it is not possible to make test scores across years comparable through test equating. Test equating presupposes the same, or at the least very similar, constructs and the same format unless it is empirically proved that there is no format effect. The reason why test equating in its rigorous sense might not be applied to achievement tests scores across years is that construct(s) to be measured might change, which is a violation of the assumption of test equating. Nevertheless, test scores from different tests across years need to be compared, which in turn requires them to be on the same metric "in a sense" as Kolen and Brennan indicate (2004).

This situation of possibly changing content coverage and test format presents peculiar challenges to those who are working on educational testing and measurement areas and also to educational researchers and policy makers. Educational researchers often use test scores as a measure of educational achievement/learning growth and policy makers make policy decisions based on test scores. If test scores across years are not comparable, evaluation of educational performance based on test scores cannot be valid and thus policy decisions based on such educational research and evaluation cannot be made correctly. Therefore, it is necessary to make sure that test scores and percentages of "being proficient" are made comparable across years even when test scores cannot be equated because of some differences in test format and contents, i.e., constructs being measured.

Because equating requires strong assumptions to be met, as described later in this work, researchers have developed different levels of linking. For example, depending on the degree of rigor, the procedures for linking test scores from different tests or test forms were classified in the past as calibration, projection, and moderation (Mislevy, 1992; Linn, 1993). Recently, Holland and Dorans (2006) classified linking procedures as predicting, scale aligning, and test equating. According to Holland and Doran's framework of linking and equating, linking of tests, i.e. test scores, which have similar constructs and similar difficulty along with similar reliability, is classified as concordance as one type of scale aligning.

In this study, test linking will refer to statistical procedures of "putting scores from two or more test forms in the same scale when the linked test scores cannot meet requirements of equating", which is very similar to an equating procedure (Kolen & Brennan, 2004). Therefore, problems related to equating design and procedures apply to test linking design and procedures. The issues related to equating are discussed to provide the context for this study comparing two different item response theory approaches to test linking[7].

With the increased application of item response theory (IRT) models in practical testing programs from test construction to test reporting, IRT equating procedures also have been increasingly developed and applied in test equating (Hambleton, 1989). Currently typical IRT equating procedures are based on unidimensional item response theory (UIRT) models. One important assumption behind UIRT equating is that multiple

[7] From now on, linking will be used to refer to test linking as defined in this study.

test forms to be equated are measuring only one construct or the same composite of multiple dimensions, which is called the "unidimensionality assumption".[8] When the assumption of unidimensionality does not hold, test data can be more adequately interpreted by analyzing with multidimensional item response theory (MIRT) models (Ackerman, 1994), which in turn suggests the need for MIRT equating.

MIRT is based on the assumption that "persons who take a test vary on a large number of cognitive demands" (Reckase & Martineau, 2004). Many researchers in educational measurement agree that many educational and psychological tests measure two or more dimensions or constructs (Ackerman, Gierl, &Walker, 2003; Briggs & Wilson, 2003). Ackerman (1994) argues that MIRT should be used to model the item-examinee interaction when test data do not meet the unidimensionality assumption. Reckase (in press) argues that the complexities of the interaction between test items and examinees (i.e., the fact that examinees use multiple skills and knowledge when they respond to items), raise the need for a model based on multiple dimensions. The fact that educational achievement tests are measuring more than one construct and therefore the probability of correctly responding the test items depends on more than one ability dimension suggests both the limitations of UIRT models and the possible usefulness of MIRT models.

One useful practical application of MIRT is the detailed analysis of content structure (Miller & Hirsch, 1992; Reckase & Martineau, 2004; Reckase 2005; Martineau, J. A., Mapuranga, R., & Ward, K., 2006). For example, Martineau, Mapuranga, and Ward (2006) in their analysis of a mathematics achievement test identified four clusters of

[8] The relationship between construct and dimension is somewhat complicated because dimension and construct do not necessarily correspond. They do sometimes, but do not when a construct can be constructed by a composite of more than one dimension.

items which are similar to each other in measuring a similar composite of multiple dimensions of mathematical ability. Martineau et al.'s work shows that classification of items into content strands by test specification does not correspond to the content structure identified by MIRT analyses. For example, one cluster identified and named as "matching data to source" using MIRT included items from three content strands. Li's work on a state mathematics achievement test (2006) also showed that classification of items by test specification does not explain the test structure identified through MIRT analysis. These findings provide supporting evidence for the argument developed in the previous section that reporting sub-scores by content strand is problematic.

Those studies document that typical mathematical achievement tests are not unidimensional. They also suggest that MIRT analyses can provide good information on what test scores really mean. If a test measures more than one ability dimension, it is necessary to analyze the data through a MIRT model. The resulting implication of these findings is that UIRT equating/linking to compare test results across years might be problematic.

1.4 The Purpose of the Study

The purposes of this study were three-fold. First, this study intended to explore what typical large-scale mathematics assessment tests are measuring and thus to provide a better understanding of what test scores really mean using MIRT dimensionality analyses of real data from a large-scale mathematics achievement assessment program. This study also used MIRT analysis to explore whether mixed format tests might provide different information on students' achievement than dichotomous tests data. Previous research conducted using a UIRT framework focused on identifying the difference in the amount of information between multiple-choice (MC) items and constructed-response (CR) items (Lukhele, Thissen, & Wainer, 1993; Donoghue, 1993). This study, on the other hand, used MIRT dimensionality analysis to explore whether and how the two types of items are different in terms of constructs they are measuring.

The second purpose of the study was to show the feasibility of conducting MIRT linking for mixed format test forms through the analyses of real data. By developing and illustrating MIRT linking procedures for mixed format test forms, this study intended to propose the practical applicability of MIRT linking to educational achievement tests of mixed format with CR items as well as MC items.

Third, this study explored the effects of selecting different linking approaches, UIRT vs. MIRT, on the proficiency classification. Specifically, the change in proficiency rate as measured by the percentage of students who were classified as proficient by UIRT linking across two years was compared with the changes that resulted from two approaches to proficiency classifications by MIRT linking. Because there is more than one ability score in MIRT, there are basically two different types of approaches to proficiency classification—conjunctive and compensatory. In the conjunctive approach, each of the scores needs to be at or above a given cut-score. In the compensatory approach, there are multiple ways of proficiency classification from the MIRT framework, depending on the weights given to each score.

For these purposes, this study selected a mathematics assessment program being administered in Ontario, Canada, as the data for the analysis because it provided a good example of large-scale mathematics assessment program of mixed format tests.

1.5 Research Questions

Specific research questions that this study intended to answer are as follows.

1. Can meaningfully interpretable dimensional structures (i.e., constructs) be identified through MIRT dimensionality analysis?

2. Do multiple-choice items and constructed-response items measure different constructs, as some researchers suggest?

3. Does the MIRT-linking approach provide an approach to sub-score reporting?

4. Are there any differences in linking results between the UIRT approach and the MIRT approaches – compensatory and conjunctive – in terms of "educational" outcomes (i.e., improvement in mathematics achievement) measured as the increase in the percentage of the students achieving a given proficiency level?

CHAPTER 2

THEORETICAL FRAMEWORK

Item response theory (IRT) is now widely being used in many testing programs because of its practical advantages over classical test theory or true score theory in solving practical measurement problems such as test equating, computer adaptive testing, optimal test design, etc. (Lord, 1980). The basic idea of IRT is to model the probability of correct response to an item as a function of ability level (i.e., ability estimate) and item characteristics (i.e., item parameters such as item discrimination, item difficulty, and pseudo guessing parameter). Item response theory can be classified either as unidimensional or multidimensional depending on the assumption of dimensionality for the response data. To provide the theoretical background for this study, unidimensional item response theory (UIRT), multidimensional item response theory (MIRT), and approaches and methods to equating/linking in the UIRT and the MIRT framework are explained in this chapter.

2.1 Unidimensional Item Response Theory

2.1.1 Unidimensional item response theory models for dichotomous data

Unidimensional item response theory (UIRT) assumes that the probability of getting an item correct can be modeled as a function of unidimensional ability. A common UIRT model (three-parameter logistic model) for dichotomous data assumes that the probability of correct response can be represented by the logistic function as follows.

$$P(U_{ij}=1\,|\,\theta_j,a_i,b_i,c_i)=c_i+(1-c_i)\frac{e^{1.7a_i(\theta_j-b_i)}}{1+e^{1.7a_i(\theta_j-b_i)}} \qquad (2.1)$$

where a_i, b_i, and c_i are item parameters characterizing an item *i*, θ_j is a person parameter indicating the level of ability being measured, and e is the mathematical constant 2.7181828..., which is the base of natural logarithms (Lord, 1980).

Specifically, the parameter b_i is usually referred to as an index of *item difficulty* and represents the point on ability scale at which an examinee has a 50% probability of getting an item correctly when the possibility of answering the item correctly by guessing is zero. Parameter c_i is called the guessing parameter or the pseudo-chance score level. It refers to the probability that a person with very low ability ($\theta=-\infty$) will answer the item correctly. When $b_i=\theta$, the probability of answering an item is .5 irrespective of the value of a if c_i is zero as the Equation 2.1 shows. If c_i is greater than zero, an examinee with the same ability (θ) as the item difficulty (b_i) has higher than .5 probability of answering the item correctly. The parameter a_i refers to item discrimination of item *i,* that is, how successful an item is in discriminating or separating examinees at an ability level, say θ_0, into different ability groups (i.e., a group of examinees with ability θ_0 or higher and the group of examinees with an ability less than θ_0 (Hambleton, 1989)). It is proportional to the slope at the point of inflection of the item characteristic curve (i.e., when $b_i=\theta$). The item discrimination parameter (a_i) theoretically can range in value from the negative infinity ($-\infty$) to the positive infinity ($+\infty$), but typical values are between 0 and 2.0.

Items with negative discrimination are discarded because it means that low ability examinees have higher probability of answering an item correctly. Both ability (θ) and difficulty (b) can range in value from negative infinity ($-\infty$) to positive infinity ($+\infty$), with typical values from -3 to + 3. The model without the guessing parameter with varying item discrimination values is called the two-parameter logistic IRT model. The Rasch model assumes that item discriminations are constant at 1 and does not include '*c*' in the model.

The following figure presents an item characteristic curve of an item with a = .9, b = .5, c = .25. The item characteristics curve (ICC) clearly shows that the probability of getting the item correct has a lower asymptote at .25 as ability decreases. When ability is equal to item difficulty at .5, the probability of getting the item is .625 instead of .5 due to the effect of the pseudo guessing parameter (c).

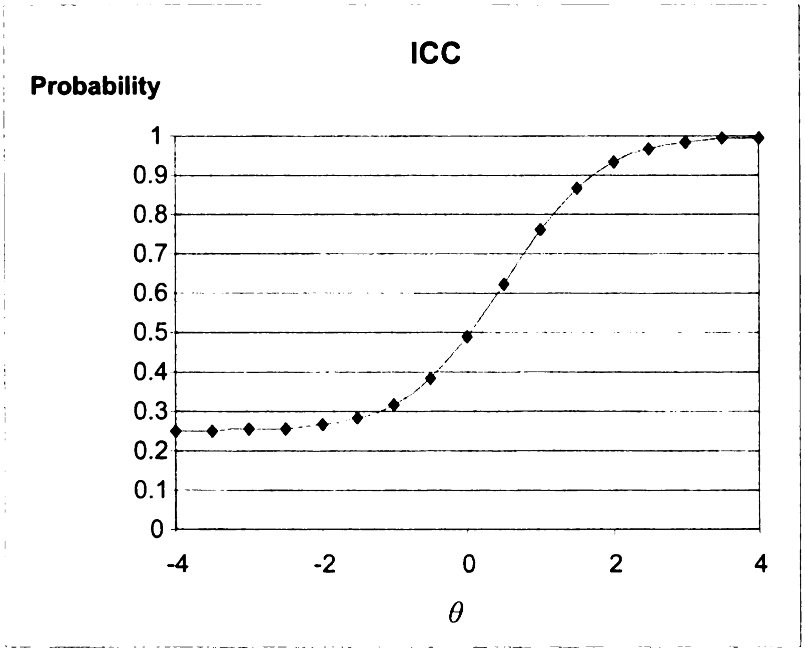


Figure 2.1 An item characteristic curve

2.1.2 Unidimensional item response theory models for polytomous data

While most test items on large scale standardized tests are dichotomous, that is, two score categories of '0' and '1', polytomous items with more than two score categories such as short-answer questions, constructed-response items, or essay questions are increasingly widely being used. To model the interaction of persons and polytomous items, a variety of models has been developed[9]. There are three models for polytomous items which have been extended to multidimensional items: the partial credit model (Masters, 1982), the generalized partial credit model (Muraki, 1992), and the graded response model (Samejima, 1969). Because this study employed the generalized partial credit (GPC) model for analysis, only the GPC model is described in detail along with brief description of the partial credit model.

The partial credit model (Master, 1982; Masters & Wright, 1997) is one of the simplest item response theory models for ordered categories data. It is appropriate for open-ended items when the complete answers (i.e., full scores) require several components to be accomplished and the overall score is given by counting the number of components accomplished. The boundaries between adjacent scores are labeled as thresholds and the model specifies the probability of a response in the categories above or below the threshold selected. In this model, item discrimination is assumed to be constant at 1 so the discrimination parameter, *a*, is not included in the model.

The generalized partial credit model (Muraki, 1992) is an extension of the partial credit model proposed by Masters (1982). The extension is the addition of the discrimination parameter, *a*, to the partial credit model. By the addition of the parameter, variation in the discriminating power of items can be modeled. The score on an item is a

[9] For a more detailed description of such models, refer to van der Linden and Hambleton, 1997.

measure of performance of the task required for the item. That is, a higher score means more accomplishment of the required task. As with the partial credit model, boundaries (i.e., thresholds) are set between adjacent scores and an examinee will have a particular probability of being on either side of a threshold. By dichotomizing the score scale for the item, the model can specify the probability of being in each of the two resulting categories. This procedure can be repeated for each threshold. After normalizing each probability at each category, *k*, within an item *i*, $P_{ikj}(\theta_j)$ so that $\sum P_{ikj}(\theta_j)=1$, the mathematical expression for the generalized partial credit model is as follows based on Muraki (1992, 1997).

$$P(u_{ij}=k\mid\theta_j)=\frac{e^{\left[\sum_{u=1}^{k}Da_i(\theta_j-b_i+d_{iu})\right]}}{\sum_{v=1}^{m_i}e^{\left[\sum_{u=1}^{v}Da_i(\theta_j-b_i+d_{iu})\right]}}, \tag{2.2}$$

where D is a scaling constant that makes the θ ability scale put in the same metric as the normal ogive model (D=1.7)

a_i is a slope parameter,

b_i is an item-location parameter,

k is the score on the item,

m_i is the total number of score categories for the item,

d_{iu} is the threshold parameter for the threshold between scores u and u-1.

The parameter b_i indicates the overall difficulty of the test item and the parameter a_i indicates the overall discrimination power of the item. The discrimination power is assumed to be the same at all thresholds, but a_i may differ across items. The threshold parameter, d_{iu}, indicates where the likelihood of responses changes from being greater for response category $k - 1$ to being greater for response category k. For estimation purposes, the sum of the d_{iu}-parameters is usually constrained to sum to 0 (Muraki, 1997).

2.2 Multidimensional Item Response Theory

Multidimensional item response theory (MIRT) models have been developed in response to the need to more accurately model the complexities of the interaction between persons and items (Reckase, 2005). It is not unusual to face test items which require more than one type of ability or hypothetical construct to solve them. One such example is a mathematical problem-solving test item which requires proficiency in both mathematical reasoning and procedural knowledge or skills. While both mathematical reasoning ability and procedural knowledge or skills can be considered as a component of broad mathematical ability, they might not be highly correlated at every level of ability. Therefore, the probability of correctly responding to the item will vary depending on various combinations of proficiency on both constructs. While unidimensional item response theory is limited in modeling more than one distinct ability dimension or hypothetical construct, multidimensional item response theory can model the relationship between more than one distinct ability dimension or hypothetical construct and examinees' different levels of proficiency across ability dimensions.

MIRT models are often classified into two types; compensatory and partially compensatory (or non-compensatory) models. The former allows low levels on one dimension to be compensated by high levels on another dimension by modeling the probability of a correct response with a linear combination of $\theta's$. In the latter, the compensation is limited because the probability of a correct response is modeled as a function of the product of probabilities for each ability dimension part. A low level on one dimension greatly constrains the overall probability of a correct response.

Research on which model fits real data better is scant and it is not possible to make any overall conclusion on the relative merits of two types of models (Reckase, in press). Because estimation with the compensatory models is relatively easy and because of the availability of computer programs such as NOHARM and TESTAFCT, however, most research and applications on MIRT have been done based on the compensatory models. [10]

2.2.1 Multidimensional item response theory model for dichotomous data

A common compensatory MIRT model for dichotomous data is given as follows.

$$P(u_{ij}=1\,|\,\vec{a}_i,d_i,\vec{\theta}_j,c_i)=c_i+(1-c_i)\frac{e^{\vec{a}_i'\vec{\theta}_j+d_i}}{1+e^{\vec{a}_i'\vec{\theta}_j+d_i}} \qquad (2.3)$$

where U_{ij} is the score (0 or 1) for person j on item i and $P\ (U_{ij}=1)$ is the probability of a correct response to items i by examinee j in a m-dimensional ability space. The form of the equation is very similar to three parameter logistic UIRT model. The difference is that the a-parameter $\vec{a}_i$ is now a vector of multiple a's which are the item discriminates along the coordinate axis and θ is also a vector that tells the location of each person on each ability dimension. As shown in the equation 2.3, d_i is a scalar parameter related to difficulty of item and c is also still a scalar. The property of the linear combination of the model can be shown by expanding the exponent part of e as follows.

[10] Partially compensatory models are not discussed because they are not directly relevant to this study. For a more thorough treatment of MIRT, refer to Reckase (in press).

$$\vec{a}_i'\vec{\theta}_j + d_i = a_{i1}\theta_{j1} + a_{i2}\theta_{j2} + \cdots + a_{im}\theta_{jm} + d_i = \sum_{\ell=1}^{m} a_{i\ell}\theta_{j\ell} + d_i \quad (2.4)$$

Equation 2.4 shows that the exponent is a linear function of the elements of $\vec{\theta}$ with the elements of the *a*-vector as slope parameters and the *d* parameter as an intercept term. When the exponent is set to a certain constant value of *k*, all theta vectors which produce *k* will give the same probability of a correct response. For example, for a test item with a_1 = .5, a_2 = 1.5, and *d* = -.75, both person A with θ_1 = 2.0 and θ_2 = -0.5 and person B with θ_1 = -3 and θ_2 = 2 will have the same probability of a correct response to the item with the value of 0.00 for the exponent of *e*. This property of the model can be shown with an equiprobable contour plot below.

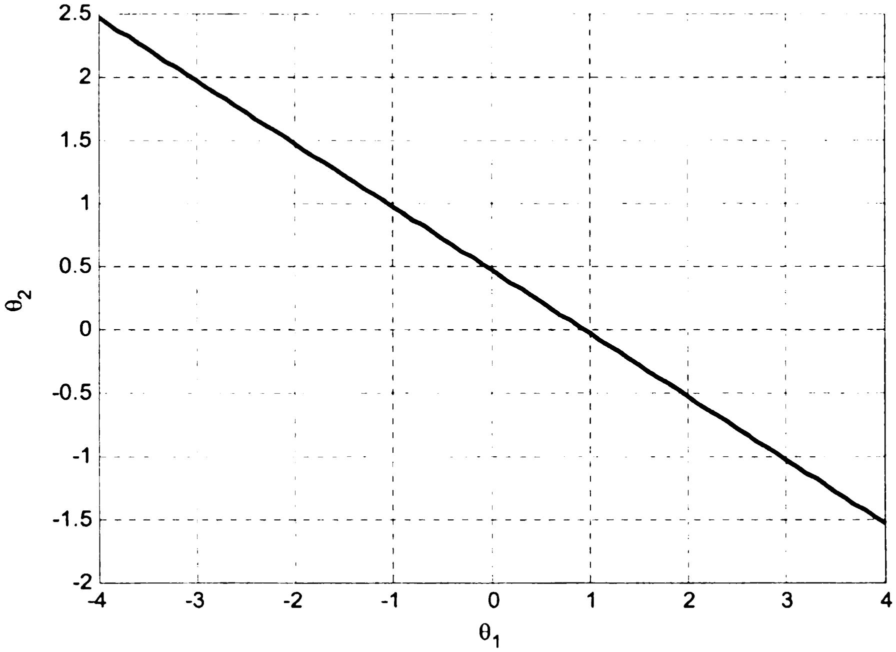


Figure 2.2 Plot of θ-vectors that yield exponents of k = 0 for a test item with parameters a_1 = .5, a_2 = 1.5, d = -.7

The compensatory property of the model is also shown in the item response surface (IRS) for an item. Item response surface shows the probability of a correct response for a particular combination of thetas for a particular item. With the same example item, the item response surface and equiprobable contour plots are shown below.

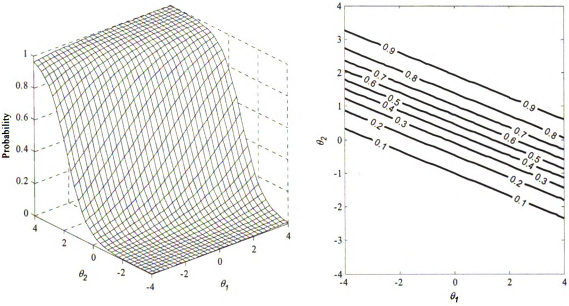


Figure 2.3 Item response surface (IRS) and equiprobable contour plots (Source: Reckase, in press)

2.2.2 Summary statistics of multidimensional item response theory

Researchers have developed a number of summary statistics that are helpful for describing items within a MIRT model framework. To describe the power of an item to separate the multidimensional space into two parts (i.e., abilities or those who respond correctly to the item and those who do not), a statistic that is analogous to the unidimensional *a*-parameter, MDISC, was developed. The overall discriminating power of an item, MDISC, is defined by the following equation (Reckase & McKinley, 1991).

$$MDISC_i = \sqrt{\sum_{k=1}^{m} a_{ik}^{\ 2}} \qquad (2.5)$$

where *m* is the number of dimensions in the ability space; and a_{ik} is an element of the $\vec{a}_i$ vector. *MDISC* is related to the slope of the item response surface in the steepest direction, that is, the direction of the best measurement of the item, and therefore analogous to the unidimensional discrimination parameter (Carlson, 1987; Reckase & McKinley, 1991). MDISC for the example item is given as 1.58 by Equation 2.5.

The graphical representation of the item vector of the same item is presented in Figure 2.4. The magnitude of MDISC is represented graphically as the length of the item vector arrow.

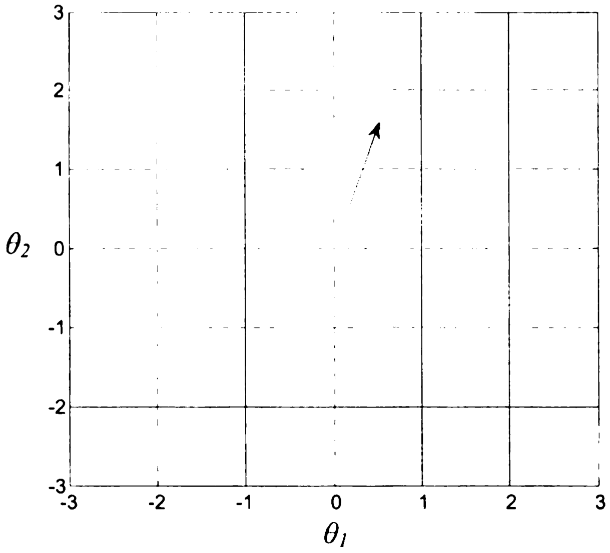


Figure 2.4 Graphical representation of MDISC by the length of the item vector arrow of the item with a_1=0.5, a_2=1.5, d=.75.

Another important statistic for describing item characteristics using MIRT is the direction of an item for the best measurement (i.e. best discrimination) in the latent ability space. The discrimination of an item is best when the angle with coordinate axis *k* is given by the following equation (Reckase & McKinley, 1991).

$$\alpha_{ik} = \arccos\left(\frac{a_{ik}}{MDISC_i}\right) \tag{2.6}$$

The angle for the best discrimination of an item defines the location (i.e., direction) of a vector of item discriminations of the item in the multidimensional ability space. For the same example item, the angle for the best discrimination, α_{i1}, is given as 72 degrees, which is the angle between the item vector and θ_1-axis. Similarity in the directions of the vector for two or more items means that they are measuring a similar combination of abilities for their best measurement in the multidimensional space. Conversely, a difference in the directions of the item vectors means that those items are measuring different combination of abilities. In Figure 2.5, the item vector of an item with the similar direction as the example item is represented as dotted-dash line. The new item has a_1=0.75, a_2=2.50, d=.75. The angle of the new item with the θ_1-axis is 73 degrees. Because of high discrimination values on both a_1 and a_2, the new item vector has a larger MDISC, as shown in the length of the new item vector.

When several items are measuring a similar combination of abilities, the direction of the best measurement for the set of items is called the reference composite. Reference composite was originally developed by Wang (1985). Wang proved (1985, 1986) that the unidimensional item response theory scale is defined by the first eigenvector of the **a'a** matrix which corresponds to the largest Eigen-value of the matrix, where **a** is the matrix of discrimination parameters for the compensatory MIRT model. The scale is a kind of average direction for the item vectors and the unidimensional θs are the values projected onto the average direction. Wang labeled the unidimensional θ that is estimated in this way as the *reference composite* for the test.

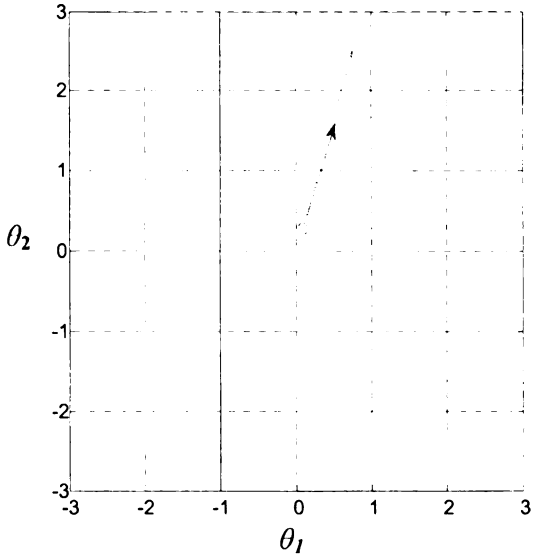


Figure 2.5 Graphical representations of two items with similar direction of best measurement.

In compensatory MIRT models, the reference composite often refers to an average direction for a set of items which are clustered together through cluster analysis. Figure 2.6 shows the arrows representing three items with similar direction of best measurement – dotted line – and the bold arrow representing the reference composite in a three dimensional θ space.

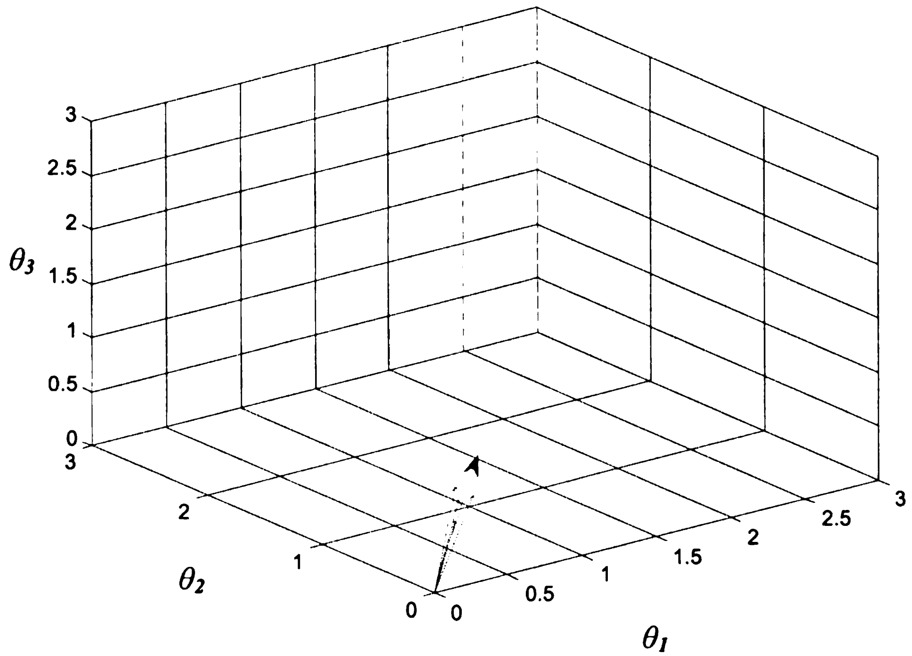


Figure 2.6 Item vectors and reference composite representing the item vectors

The angle between the reference composite and the coordinate axes can be determined by taking the arccosine of the elements of the eigenvector. In this example, the reference composite has an angle of approximately 46° with the θ_1 axis, 46° with the θ_2 axis, and 78° with the θ_3 axis.

Multidimensional difficulty (MDIFF) is graphically represented as the distance of the point of maximum slope from the origin (Figure 2.7) and its mathematical expression is as follows (Reckase, 1985).

$$\text{MDIFF}_i = \frac{-d_i}{MDISC_i} \tag{2.7}$$

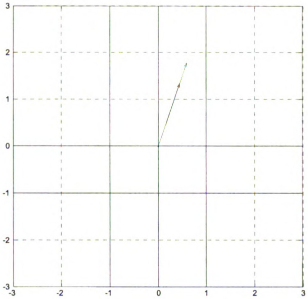


Figure 2.7 Item vectors with MDIFF of zero and a positive value

In Figure 2.7, the dotted arrow is an item vector which has MDIFF of zero, i.e., beginning on the origin and the distance of the item vector with a solid line from the origin indicates the magnitude of MDIFF for the item. The interpretation of MDIFF is similar to that of the difficulty parameter in the unidimensional model. The magnitude of the MDIFF of an item is the distance of the item from the origin. If the starting point of the item vector is located in quadrant III, it has a negative value.

In summary, the full description of an item in a multidimensional space is given by the direction of best measurement ($\alpha_{ik} = \arccos(\frac{a_{ik}}{MDISC_i})$), the discrimination in that direction (MDISC), and the distance of the point of the best discrimination from the origin (i.e., multidimensional difficulty (MDIFF)) (Reckase, 2005). In a multidimensional space, item and test characteristics are represented as a surface instead of a curve, as in UIRT.

2.2.3 Multidimensional item response theory model(s) for polytomous data

For polytomous data, several MIRT models have been proposed. Reckase (in press) describes the MIRT extensions of the generalized partial credit, partial credit, and graded response models. Reckase indicates that there is no MIRT version of the partially compensatory model proposed for the polytomous case.

In the multidimensional extension of the generalized partial credit model (MGPC), ability estimates and item discrimination estimates are represented as vectors to describe separate values for each ability dimension, but the threshold parameter is assumed to be constant across ability dimensions at each score category. Because this study employed the MGPC model for analysis, only the MGPC model is described.

Reckase's presentation of the model below is a slight variation of the model given in Yao and Schwarz (2006).

$$P(u_{ij} = k \mid \boldsymbol{\theta}_j) = \frac{e^{k\mathbf{a}_i\boldsymbol{\theta}_j' - \sum_{u=0}^{k} \beta_{iu}}}{\sum_{v=0}^{K_i} e^{v\mathbf{a}_i\boldsymbol{\theta}_j' - \sum_{u=0}^{v} \beta_{iu}}} \tag{2.8}$$

where β_{iu} is the threshold parameter for score category u,
β_{i0} is defined to be 0,
and all other symbols have their previously defined meaning.

The item response surfaces for the MGPC model for test data which can be represented in a space with two coordinate dimensions are given in Figure 2.8 as presented by Reckase (in press). The test item represented here has scores from 0 to 3. The item parameters for the model are $\mathbf{a_i}$ = [1.2 .7] and β_{iu} = 0, -2.5, -1.5, .5.

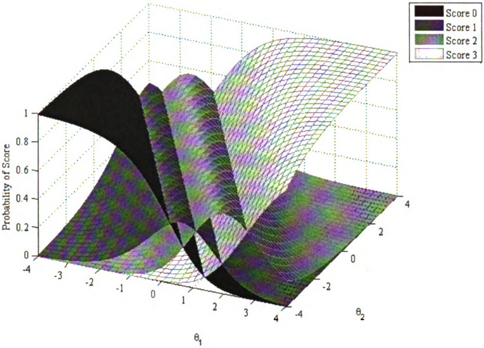


Figure 2.8 The item response surfaces for the MGPC model (Source: Reckase, in press)

The intersections between the surfaces for adjacent score categories are represented as a straight line in the **θ**-plane. In general, the line consists of the set of points in the **θ**-plane where the probabilities of obtaining the adjacent scores are the same (Reckase, in press). Reckase indicates that MGPC can be simplified in a number of ways so that they can have the special properties of the Rasch model (i.e., observable sufficient statistics). One such model has been presented by Kelderman and Rijkes (1994).

2.3 Linking and Equating

Equating can be considered as the most rigorous linking procedure. Because equating presupposes scale linking in both item parameters and ability estimates, discussion of the theory and method of equating is directly applicable to linking design and method. For this reason, theory and method of equating are presented with special attention to the relationship between scale linking and equating.

2.3.1 Unidimensional item response theory and test linking/equating

Two important characteristics of item response theory compared to classical test theory are invariance of item parameters and scale indeterminacy. Item parameters in item response theory remain the same irrespective of either the ability distribution of the group who take the tests or the overall difficulty level of the tests taken. That is, in theory, item parameters are invariant across groups (Lord, 1980). On the other hand, in classical test theory, item difficulty and item discrimination indices to evaluate the quality of items and thus of tests vary depending on the groups or overall test difficulty. In classical test theory, for example, if a group of test takers are homogeneous in their ability measured, most items on a test can be very easy when a high ability group is taking the test or very difficult if a very low ability group is taking the test. In this situation, the test items do not discriminate either of the two groups of examinees well. If test takers are heterogeneous, test items will discriminate examinees well, and difficulty of items can decrease or increase.

However, item difficulty in item response theory is assumed to be the same for persons with the same ability across different groups in the population. That is, the probability of correctly responding to an item is the same for persons with the same ability irrespective of which tests they take if the tests are designed to measure the same latent trait, i.e., construct, or the same combination of them.

Also, while examinees' true scores in classical test theory will increase when they take easier tests, ability estimated through IRT models remains the same whatever combination of item sets they respond to. In one- and two-parameter IRT models, for example, an item difficulty parameter (b) is the point on the ability (θ) scale at which an examinee has a 50% probability of answering the item correctly. Therefore, when test items become easier, the probability of correct response to the items for the same ability examinees will increase, but higher probability of correct response, that is, increased test scores, does not mean increased ability—invariance of ability parameters. In IRT, population invariance of item parameters is assumed to hold if data fit the model. However, if the origin and unit of the ability scale change, the invariance assumption cannot hold unless corresponding changes are made to the item parameters. This raises the need to fix the origin and unit of the ability scale.

Item response function in Equation 2.1 is expressed as a function of $a_i(\theta - b_i)$, where i refers to an item. If we add the same constant to each θ and at the same time to every b_i, the quantity of $a_i(\theta - b_i)$ remains the same and so does the item response function, i.e. $P_i(\theta)$. This shows that the choice of origin for ability scale is purely arbitrary (Lord, 1980). In the same fashion, the choice of unit for measuring ability is also purely arbitrary. Typically, the statistical procedures used by IRT computer programs calibrate parameters so that mean and standard deviation of ability estimates are 0 and 1.

This means that the scales of the parameters set up to estimate them are applicable to only the data analyzed. That is, it is possible that two persons with the same ability have different ability scores when they take different sets of items because item parameter estimates of the same items from different calibrations are not the same. However, this is because abilities were measured using different metrics, so the different values for the same items need to be put on the same metric to make test scores from different tests or test forms comparable.

The procedure for putting item parameters on the same metric is sometimes called scaling, linking, or calibrating. In this study, the procedures to put different estimates of items parameters of the same items across different test forms on the same metric has been referred to as scale linking. The procedure of estimating item parameters will be called "calibrating". Scaling is used to refer to the process of associating numbers with the performance of test takers as Petersen, Kolen and Hoover define (1989). Equating refers to the whole procedure of making ability scores or reported scores from different test forms comparable, as already defined.

Test equating is conducted to produce exchangeable scores on different test forms which are assumed to be designed to measure the same construct or constructs, typically by the same test specification. The two characteristics of IRT discussed above, i.e., invariance of IRT parameters and scale indeterminacy, define the alternatives for IRT equating, depending on the equating design.

2.3.2 Equating design and methods

There are several designs for test equating. One of the commonly used designs for IRT equating is non-equivalent group common item design. Especially in educational testing situations, groups who take different test forms cannot be assumed as equivalent. In fact, equivalence of groups cannot be assumed unless groups are selected completely at random. For this reason, and because it is easy to administer common items on different test forms, non-equivalent group common item design is one of the mostly widely used methods in IRT test equating. When the non-equivalent group common item design is used, there are basically three possible ways to approach equating.

2.3.2.1 Concurrent calibration

In concurrent calibration equating design, all examinee response data from separate test administrations using different test forms with some common items are calibrated in a single run. When the data are calibrated concurrently, item parameters estimates and ability estimates are already on the same metric, so it is possible to compare ability scores directly. In this case, scale transformation, i.e. linking, is not necessary.

2.3.2.2 Fixed common item parameters (FCIP)

When item parameter estimates for the common items are available from one test data set or calibration, those parameters can be used to put the item parameters from the other tests on the same metric as the metric for the previous test data. This is done by fixing the item parameter estimates for the common items as the values calibrated from the previous test data set when the test data for the later administration are calibrated. This places all the item parameter estimates for the later test data set on the same metric as the previous one. Jodoin, Keller, and Swaminathan (2003) call this approach as "fixed common item parameter calibration equating design" (FCIP). Their research suggested that selection of the linking calibration method makes a difference in proficiency classification, but they cannot say which method is more accurate because their findings are based on the analysis of empirical data. In practical educational testing situations, the contemporary test results are compared with the previous ones, so the previous tests are used as the base test. When FCIP linking method is used, there is no further linking/equating procedure for item parameters or ability estimates because they are already on the same metric.

2.3.2.3 Linking after separate calibrations for dichotomous data

When response data from different test forms, i.e., different test administration, are calibrated separately, it is necessary to take procedures to put the item parameters on the same metric because the origin and unit of item parameters from the separate calibrations are different, as already discussed. Linking or scale transformation to put the item parameters on the same metric can be conducted in several ways, but they can be classified in two categories, moment methods or characteristic curve (CC) methods (Kim, 2004). There are two commonly used moment methods—mean/sigma and mean/mean. Two common CC methods are the Haebara approach and the Stocking and Lord approach (Kolen & Brennan, 2004).

When there are different sets of item parameters for the common items from the separate calibrations, they can be linearly transformed using the proper formula because the same items should have the same parameters, as the assumption of population invariance of parameters suggests. A simple example of a linear transformation is the F=32+C*5/9, which is used for restating a temperature measured on the Celsius scale to one expressed in Fahrenheit.

In the case of 3PL logistic model, the relationship between two separate scales from separate calibrations is as follows. (Kolen & Brennan, 2004)

$$\theta_{Ji} = A\theta_{Ii} + B \qquad (2.9)$$

$$a_{Ji} = \frac{a_{Ji}}{A} \qquad (2.10)$$

$$b_{Ji} = Ab_{Ij} + B \qquad (2.11)$$

And $c_{Jj} = c_{Ij}$

where a_{Jj}, b_{Jj}, and c_{Jj} are the item parameters for item j on scale J and a_{Ij}, b_{Ij}, and c_{Ij} are the item parameters for item j on scale I. The pseudo guessing parameter or lower asymptote parameter is independent of scale transformation. Equations 2.9 to 2.11 express the relationship between scales by two abilities and two items. The relationship in terms of groups of items or people can be expressed as follows.

$$A = \frac{\sigma(b_J)}{\sigma(b_I)} \qquad (2.12a)$$

$$= \frac{\mu(a_J)}{\mu(a_I)} \qquad (2.12b)$$

$$= \frac{\sigma(\theta_J)}{\sigma(\theta_I)} \qquad (2.12c)$$

$$B = \mu(b_J) - A\mu(b_I), \text{ and} \qquad (2.13a)$$

$$= \mu(\theta_J) - A\mu(\theta_I) \qquad (2.13b)$$

As shown by the equations 2.12a to 2.13b, the constants, *A* and *B*, for scale transformation can be computed from the relationship between IRT parameters of two scales and using these constants, *A* and *B*, scale transformation can be done. There are several methods for scale transformation or linking, which are described below.

2.3.2.3.1. The mean/mean and mean/sigma methods

In mean/mean method, which was originally described by Loyd and Hoover (1980), the mean of the *a*-parameter estimates for the common items is used in place of the parameters in equation (2.12b) to estimate the *A*-constant. Then, the mean of the *b*-parameter estimates of the common items is used in place of the parameters in equation (2.13a) to estimate the *B*-constant (Kolen & Brennan, 2004).

In mean/sigma method, which was originally described by Marco (1977), the means and standard deviations of the *b*-parameter estimates from the common items is used in place of the parameters in equations (2.12a) and (2.13b) (Kolen & Brennan, 2004).

2.3.2.3.2. Characteristic curve transformation methods

The procedures developed by Stocking and Lord (1983) and Haebara (1980) are characteristic curve transformation methods. Characteristic curve transformation methods involve finding the slope (*A*) and intercept (*B*) of a linear scale transformation function so as to minimize the discrepancy between characteristic curves using parameter estimates on the target scale and characteristic curves using parameter estimates on the transformed scale by this linear function. (Kim & Hanson, 2002)

a. Haebara approach

The function used by Haebara (1980) to express the difference between two item characteristic curves is the sum of the squared difference between the item characteristic curves for each item for examinees of a particular ability. The difference between each item characteristic curve on the two scales is squared and summed (Kolen & Brennan, 2004).

b. Stocking and Lord approach

In contrast to the Haebara approach, the Stocking and Lord approach uses the sum of the squared differences for all common items as criterion. In this approach, the transformation constants, A and B, are chosen to minimize the weighted sum of squared distances between two test characteristics curves from the common items on two test forms.

2.3.2.4 Polytomous IRT models and equating of mixed format tests

Linking procedures for polytomous items have been developed by extending the linking procedures developed under dichotomous IRT models. For example, Baker (1992, 1993) extended the Stocking-Lord procedure to Samejima's graded response model (GRM) and Cohen and Kim (1998) extended the mean/mean and mean/sigma procedures to the GRM. For Muraki's (1992) generalized partial credit model (GPCM), formulas to extend minimum chi-square, Haebara, and Stocking-Lord procedures were proposed by Hattori (1998)[11].

When test data consist of both dichotomous and polytomous items, the first condition to be considered in linking/equating mixed format test forms is the dimensionality of test structure. If the tests can be considered to meet a unidimensionality assumption, all three equating design approaches discussed above can be applied. Under the three approaches, there are two additional options. That is, test data can be calibrated separately by format or jointly across different formats. When mixed format data can be analyzed under UIRT models, simultaneous calibration is more useful when a score scale as a summed score from each format is required and calibration by format will be useful

[11] Recited from Kim, S. (2004)

when scores for each format are reported separately in addition to a summed score (Kim, 2004).

As explained already, if concurrent calibrations or item parameter estimates from the base test forms can be used as fixed for the calibration of data to be equated, no further linking procedures are required when ability scores are compared. If separate calibrations are used for mixed format tests, two separate sets of item parameter estimates should be placed on a common scale. For this, Li, Lissitz, and Yang (1999) proposed an extended version of the Stocking-Lord linking procedures for mixed format tests, for which the three-parameter logistic model and the GPCM were used. Tate (2000) also presented linking procedures for mixed format tests for multiple-choice items and constructed-response items by extending the mean/sigma and Stocking-Lord linking procedures.

2.3.2.5 Comparison of linking procedures

For dichotomous IRT models, research comparing the characteristic curve methods and moments methods seems to agree on the better performance of the former (Kolen & Brennan, 2004). Kim (2004) reports the same finding for mixed format test equating. When scale transformation methods are compared to the concurrent calibration method, the research findings are more favorable for concurrent calibration (Kim & Cohen, 1998; Kim, 2004). Kolen and Brennan (2004) indicate that previous studies suggest that concurrent calibration method might be less robust to the violation of unidimensionality assumption. They recommend separate estimations using the test characteristic curve methods.

2.3.3 MIRT linking and reference composites

In addition to two types of indeterminacy (i.e., the origin and the unit discussed for UIRT) there is the third type of indeterminacy in MIRT – rotational indeterminacy. The orientation of the axes of the coordinate system defined by each calibration is different. The goal of MIRT linking/equating procedures is to find transformations that will convert a set of item parameter estimates from one data set to the estimates from another data set. The comparable transformations for abilities can be determined using the transformation matrix for the common items.

The first attempt to deal with multidimensionality in IRT linking was made by Hirsch (1988, 1989). Later, other researchers have worked on MIRT linking methods (Li & Lissitz, 2000; Oshima, Davey, & Lee, 2000; Min, 2003). Li and Lissitz (2000) resolve the three indeterminacy problems by using a translation vector **m**, a scalar dilation parameter *k,* and orthogonal Procrustes rotation matrix **T**, respectively. Identifying a limitation of the scalar dilation parameter proposed by Li and Lissitz, Min (2003) developed a MIRT equating procedure using a diagonal dilation matrix that allows for differential dilation/compression of the scales of the various dimensions (Reckase & Martineau, 2004). Reckase and Martineau (2004) found that Min's approach to MIRT equating brings an infeasible burden of computation when dimensionality is high. To correct this weakness of Min's approach, they proposed using non-orthogonal Procrustes transformation.

Reckase (in press) proposed a more general approach than previously published methods by extending the methods based on work by Martineau (2004) and Reckase and Martineau (2004). Reckase proposes transforming the calibrations of items to be on the same scale using reference composites consisting of items measuring same constructs or sub-constructs. The first step proposed by Reckase is to determine the rotation matrix for the coordinate axes. The rotation matrix is given by the following equation.

$$\mathbf{Rot} = (\mathbf{a}'_a\, \mathbf{a}_a)^{-1}\, \mathbf{a}'_a\, \mathbf{a}_b \qquad (2.14)$$

where $\mathbf{a}_b$ is the n x m matrix of base form discrimination parameters that are the target for the transformation, $\mathbf{a}_a$ is the n x m matrix of discrimination parameters for the same items on the alternate form, and **Rot** is the m x m rotation matrix for the discrimination parameters.

Then, the *d*-parameters from the alternate form are transformed to the metric of the base form using the following equation.

$$\mathbf{Trans} = \mathbf{a}_a\, (\mathbf{a}'_b\, \mathbf{a}_b)^{-1}\, \mathbf{a}'_b\, (\mathbf{d}_a - \mathbf{d}_b) \qquad (2.15)$$

Where $\mathbf{d}_b$ is the n x 1 vector of d parameter for the base form, $\mathbf{d}_a$ is the n x 1 vector of *d*-parameters for the alternate form, and **Trans** is the n x 1 transformation vector for the *d*-parameters.

Last, the transformation of the estimates of θ from the alternate form to the base form metric is given by the equation (2.16).

$$\widehat{\theta}'_b = \mathbf{Rot}^{-1} \boldsymbol{\theta}'_a\, (\mathbf{a}'_b\, \mathbf{a}_b)^{-1}\, \mathbf{a}'_b\, (\mathbf{d}_a - \mathbf{d}_b) \qquad (2.16)$$

where $\boldsymbol{\theta}'_a$ is a 1 x m vector matrix of estimates from the alternate form calibration and $\widehat{\theta}'_b$ is the 1 x m parameter estimate vector after transformation to the coordinate system from the base form.

The point is that the probability of answering an item correctly should be the same before and after transformation (i.e., linking) through translating and rotation. When

$$\hat{a}_b = a_a T$$
$$\hat{d}_b = d_a + aTm$$
$$\hat{\theta}_b = T^{-1}\theta'_a - m$$
$$m = (\hat{a}'_b \hat{a}_b)^{-1} \hat{a}_b'(d_b - d_a)$$

the probability after the transformation remains the same, as the following equations show.

$$\begin{aligned} p &= \hat{a}_b \hat{\theta}_b' + \hat{d}_b \\ &= aT(T^{-1}\theta' - m) + d_a + aTm \\ &= a\theta' - aTm + d_a + aTm \\ &= a\theta' + d_a \end{aligned}$$

Reckase indicates that the transformations using the equations presented here can be used for multidimensional generalizations of horizontal equating. The procedures described here are useful when linking two different test results after separate calibrations.

In MIRT linking, the FCIP linking method which was described in the UIRT linking section has not been tried. One reason might be the absence of a program which can perform the FCIP MIRT linking. Another reason is that MIRT linking has been applied mostly in vertical linking context (Martineau, 2004; Yon, 2006; Li, 2006). In this study there was no need for scale transformation because item parameters were put on the same metric through the FCIP linking. It is possible to apply the procedure described above after putting the common items on the same metric through the FCIP linking or the concurrent calibration.

Every linking method has its own advantages and shortcomings. The FCIP method is expected to produce relatively small measurement error because it skips one step in the estimation process. Jordin, Keller, and Swaminathan (2003) suggested that the FCIP method is relatively more accurate, through on the basis of a study of empirical data, but they could not provide any information on relative performance of different approaches in terms of recovery of true parameters because they did not conduct a simulation study. A recent study reports little difference between concurrent calibration and FCIP linking/equating method in terms of item recovery based on simulated data (Taherbhai & Seo, 2007).

CHAPTER 3

RESEARCH DESIGN, DATA, AND METHODS

This chapter describes the research design, data, and methods used for this study. The research design is briefly described, followed by the description of the data. Then each procedure for the study described in the research design is explained in detail.

3.1 Research Design Overview

The data for this study were from the grade 6 assessment tests in mathematics for the province of Ontario, Canada. The procedures in this study can be broadly classified into three parts; 1) UIRT linking to compare proficiency rates based on unidimensional ability between 2005 and 2006; 2) MIRT linking to compare proficiency rates based on multidimensional abilities between 2005 and 2006; 3) comparing proficiency rate change from the UIRT linking approach with two proficiency rate changes from the MIRT linking approach between 2005 and 2006.

In the first part of the study, the 2005 test result was linked to the 2006 test result through the fixed common item parameter (FCIP) UIRT linking method. The procedure and rationale for the selection of the linking method are described in detail below. By linking two years' test data, the change in the proficiency rate (i.e., the percentage of students who are at or above proficiency between two years) was explored.

The second part of the study, MIRT linking, consisted of three stages. In the first stage, MIRT linking for two year's test data was conducted using the same FCIP method. The multidimensional ability structure of the data was explored and 2005 test data were put on the same scale as the 2006 test data through FCIP linking. In the second stage, MIRT cluster analyses of the data from the two tests were conducted. In addition, a review of the content of test items was performed by the item review committee[12]. In this step of the study, three "reference composites" were identified and they were interpreted as constructs being measured by the tests. The third stage of the MIRT linking consisted of projecting multidimensional ability onto the reference composites. Ability scores estimated on multiple dimensions were projected on each reference composite. This approach resulted in construct-level ability scores that could be used for proficiency classification.

The third part of the study consisted of comparing the proficiency rate change from UIRT linking with the two proficiency rate results obtained from the two different approaches to MIRT proficiency classification, compensatory and conjunctive.

3.2 Data and Samples

This study analyzed 2005 and 2006 grade 6 mathematics test data. In 1995, Ontario province in Canada enacted legislation to establish the Educational Quality and Accountability Office (EQAO) for the purpose of providing "accurate, objective and clear information about student achievement and the quality of publicly funded education in Ontario"[13]. The major role of EQAO was to establish and conduct a provincial testing program for students in Ontario's English or French language schools. The main purpose of the testing program is "to provide accurate and valid data about student performance." EQAO administered its first annual assessment in the 2000-2001

[12] The item review committee is explained in Methods section below.

[13] Retrieved on Feb. 2007 from http://www.eqao.com/AboutEQAO/GeneralQuestions.aspx?Lang=E

academic year. This study analyzed only tests for English-language schools, excluding tests administered in the French-language schools.

The assessment framework was developed by EQAO based on the province-wide mathematics curriculum framework, the Ontario Curriculum (Jackson, 2007).[14] Overall and specific expectations are organized into five content strands—Number Sense and Numeration, Measurement, Geometry and Spatial Sense, Patterning and Algebra, and Data Management and Probability. This mathematics content area classification corresponds to the five content strand classification in *Principles and Standards for School Mathematics* (PSSM) (NCTM, 2000) and many other U.S. state mathematics curriculum frameworks. The assessment framework also specifies learning expectations in mathematical processes. They are problem solving, reasoning and proving, reflecting, selecting tools and computational strategies, connecting, representing, and communicating.

The 2005 test consists of 42 items; 32 multiple-choice items, 5 short-answer items[15], and 5 open-response items[16]. The 2006 test consists of 36 items; 28 multiple-choice items and 8 open-response test items. Total number of the students was 132,021 for the 2005 test and 136,653 for the 2006 test. While the two tests were developed based on a similar assessment framework, test specifications were different, as shown by the different formats of the two tests.

[14] Ontario grade 6 mathematics curriculum can be downloaded from http://www.edu.gov.on.ca/eng/curriculum/elementary/math6ex/

[15] Short-answer items are open-ended items requesting short answers. These items are scored "0" or "1".

[16] Open-response items are open-ended items requesting extended answers. This type of items is often called as constructed-response items in the United States. Therefore, open-response items in Ontario assessments are described as constructed-response items later in this chapter.

For the purposes of linking the two years' test data and field-testing items for the next year's test, five field-test items—four multiple-choice items and one open-response—were administered to each student when they took the 2005 test. Of the 36 operational items in the 2006 test, 34 items were from the field-tested items in 2005. Not all field-tested items were used in the 2006 operational test. The 34 items were used as common anchor items for linking. Two open-response items in the 2006 test were translated from French operational test items for French-speaking schools.

Field-test items were administered using matrix sampling to minimize the number of items administered to each student. The relationship between this sampling method and procedures used in this study is described below. Because some field-tested items were excluded from the 2006 test because of low psychometric quality, each student in 2005 contributed zero to four items to the response data used as anchor items while all students had five field-tested items. There was no anchor item administered to all of the 2005 students, which will be illustrated in Table 3.1. This study was constrained by this real data structure as described in detail below.

3.3 Procedures of the study

This part describes procedures for the study, the decisions made to accommodate the real data structure, and the reasons and the rationales for changes made during the process of conducting the research whenever necessary.

3.3.1. Selection of linking method and computer software programs

First, for the purpose of comparison, it was decided that the same linking procedure should be applied to both UIRT and MIRT linking. The linking method originally planned for this research study was scale-linking using common items after separate calibrations. For UIRT linking, PARSCALE 4 (Muraki & Bock, 2002), an IRT computer software program for mixed format tests, was selected because many testing companies are adopting the program for IRT analyses of mixed format test data. After separate calibrations, two test scales were planned to be linked using the STUIRT program[17] (Kim & Kolen, 2004a) and the ability scores were designed to be linked through the POLYEQUATE program[18] (Kolen, 2004a). For MIRT linking, it was planned to adopt an oblique Procrustes transformation method after separate MIRT calibrations using the BMIRT program, a program for MIRT calibration of mixed format tests. Because BMIRT uses the Markov Chain Monte Carlo (MCMC) procedure for estimation and thus calibrating data takes a long time, only five random sample data sets for each year were used for replications.

The sample size of at least 5000 was considered to be necessary for stable estimation of MIRT parameters. Working on five random samples of the full data (RSD) is equivalent to conducting five replications of linking of the 2005 test results with the 2006 test results.

The original plan for this linking study as described above was made before the real data were available. Close examination of the real data suggested that changing original study plan was necessary. The data structure constrained the choice of linking method and the size of the RSD sets. Because of the matrix sampling design for the field test items of the 2005 test, the number of cases for some score categories on the

[17] STUIRT is a computer program for implementing the four IRT scale transformation methods extended by Kim and Lee (2004) for use with mixed-format tests as well as single-format tests (Kim and Kolen, 2004b).

[18] Program POLYEQUATE is a Fortran 77 program that conducts item response theory (IRT) true and observed score equating using dichotomous and polytomous IRT models. (Kolen, M. J., 2004b)

constructed-response (CR) items was too small for stable calibration. The smallest number of respondents for such items was 32 for data sets with 5000 cases. This means there were too many purposeful missing values due to data collection design—more than 90% missing values for anchor items. The calibration results from PARSCALE for the 2005 RSD sets were expected to be unstable. For this reason, a decision was made to adopt the FCIP linking method instead of scale transformation after separate calibration. FCIP linking was judged as reasonable and practical because it was being used for operational equating by EQAO. Also, MIRT linking through FCIP would be possible using the BMIRTanchor program[19].

After the change of linking method was made, the sample size of the RSDs was also changed. Because of small number of respondents to constructed-response items, there were some response categories with no responses. The PARSCALE program does not calibrate polytomous data with response categories with zero frequencies. One solution often taken in this situation is merging or collapsing two or more response categories. This approach was not taken because the research question of whether multiple-choice items and constructed-response items are measuring the same construct would not be answered appropriately. Instead of merging response categories, it was decided to analyze larger size RSDs. Practical considerations of the time required to run BMIRT and BMIRTanchor constrained the size of RSDs as well. It was expected that 10,000 sample data sets would allow running PARSCALE without spending too much time for MIRT calibration using BMIRT and for MIRT linking using BMIRTanchor.

[19] There were 8 different versions of BMIRT program which were designed to be used for different purposes when this study was conducted. In this study, BMIRT28 for MIRT calibration and BMIRTanchor for FCIP MIRT linking were used.

From the entire response data of two year's tests, 13 RSD sets of 10,000 were sampled for each test.

3.3.2 Selecting random samples data sets for the study

For both UIRT and MIRT linking, 13 random samples data (RSD) sets of 10,000 were sampled without replacement for each year. In the case of 2006 test data, RSD sets were selected without any difficulty using MATLAB. However, selecting RSD sets of the 2005 data with anchor items was a very complicated process because of the matrix sampling design for the testing program. To make sure that each RSD set includes the same proportion of the same anchor group, i.e., a group of students who took the same anchor items, the entire test data had to be divided into distinct groups with the same set of anchor items. The classification of the test data into the same anchor items groups was made using the SAS program. Using SAS, the data structure was explored based on the missing data pattern and 22 distinct anchor groups with different combinations of anchor items were identified. Each group had zero to four anchor items. To decide the size of each anchor group within each RSD set, the total number of each anchor group was divided into 13 so that adding all samples from 22 distinct anchor groups produced data sets of 10,000 each. From each anchor group, 13 data sets were selected using MATLAB. Then, 13 RSD sets of 10,000 were created by combining the 22 anchor group samples. From the 13 RSD sets from year 2005, only 6 RSD sets could be used for PARSCALE calibration because other RSD sets had empty response categories. Among the six 2005 RSD sets, five RSD sets were used for this study.

The field-test items for the 22 distinct groups and the number of samples from each anchor group per RSD are reported in Table 3.1. Among the 34 anchor items, 25 items were administered to one group and 9 items were administered to two groups. The combinations of the field test items for the 22 distinct groups are reported in Table 3.2. In Table 3.2, column *G* indicates the distinct anchor group and column *I* indicates the number of items administered to each anchor group. The next 34 columns are for the 34 anchor items. The table shows how matrix sampling works. An X inside a cell denotes that that anchor group had that item and a blank space indicates the item was not presented. Items not presented are intentionally missing data. While missing was a result of sampling design, it still presents a problem for estimation.

Table 3.1 Number of cases for each anchor group in each RSD

Anchor Group 1	83
Anchor Group 2	68
Anchor Group 3	87
Anchor Group 4	90
Anchor Group 5	68
Anchor Group 6	64
Anchor Group 7	657
Anchor Group 8	641
Anchor Group 9	566
Anchor Group 10	543
Anchor Group 11	558
Anchor Group 12	630
Anchor Group 13	620
Anchor Group 14	624
Anchor Group 15	630
Anchor Group 16	621
Anchor Group 17	628
Anchor Group 18	517
Anchor Group 19	528
Anchor Group 20	614
Anchor Group 21	551
Anchor Group 22	612
RSD	10000

Table 3.2 Anchor groups and anchor items field-tested to the groups

G	*I*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1	1	x																																	
2	2		x								x																								
3	2			x																									x						
4	3				x																						x	x							
5	3					x							x	x																					
6	4						x																										x	x	x
7	1							x																											
8	2								x	x																									
9	1										x																								
10	1											x																							
11	2												x	x																					
12	2														x	x																			
13	3																x	x	x																
14	3																			x	x	x													
15	2																						x	x											
16	1																								x										
17	1																									x									
18	2																										x	x							
19	1																												x						
20	3																													x	x	x			
21	3																																x	x	x
22	0																																		

3.3.3. UIRT linking

As explained in the previous chapter, item parameters for the common anchor items are fixed as pre-calibrated estimates from a base or reference data set and all the other items of the data set to be linked are estimated on the same scale as the fixed common items so that ability scores from two data sets are put on the same scale.

For this study, the 2006 test data were treated as the reference data and the 2005 test data were linked to the 2006 test data. The 2006 test data were expected to produce much more stable item parameters for the 34 common anchor items compared to the 2005 test data because the 34 common items have responses from 10,000 examinees for the 2006 test, but from only 68 to 657 samples for the 2005 test. RSD sets for the 2005 test data were matched with five RSD sets for the 2006 test so that 5 distinct pairs of RSD sets were made. Each of the paired RSD sets was linked using FCIP linking method.

The specific procedures taken for UIRT linking are as follows. First, RSD sets generated through MATLAB were saved as files for PARSCALE calibration. Then, five random samples data sets of 10,000 each of the 2006 test were calibrated and abilities were estimated. Third, RSD501A[20] was calibrated so that item parameter file produced could be used for FCIP linking. For FCIP linking, three files are needed for calibration of each 2005 RSD set; data file, item parameter file, and NF file to indicate "not-presented" items. To calibrate each 2005 RSD set as fixed on the 2006 scale, five 2005 RSD sets were calibrated with fixed item parameters for the common items as estimated from calibrating 2006 RSD sets.

[20] In this study, there are three types of RSD sets; the 2006 test, the 2005 test, and the 2005 test with anchor items or common items. To distinguish the 2005 test and the 2005 test with anchor items, the latter was labeled as RSD501A to RSD505A. After linking of the RSD2005A with the 2006 matched RSD sets, the 2005 RSD sets were labeled as RSD501 to RSD505 without A.

By fixing item parameters for the field test items for the 2005 test data, the resulting calibrated item parameters for the 2005 test data are on the same metric as the 2006 test data, so estimated abilities for the 2005 test data are also on the same scale as the one for the 2006 test calibration. This means that abilities estimated for both years were on the same metric so they could become comparable.

For this study, 80% of 2005 students were arbitrarily selected as the proportion assumed to have achieved the proficiency level or higher when applying a cut score produced through a standard setting procedure. The cut score, which was expressed as a θ value, for the proficiency level in 2005 was obtained at the 20th percentile. This cut-point score was applied to the 2006 test data to get a proficiency rate in 2006 for each RSD set. This cut score was applied to obtain the percentage of the students who passed the cut score for proficiency level in 2006 and thus to calculate the change of proficiency rates between 2005 and 2006. The same procedure was applied to each RSD set and the average of proficiency change was calculated to be compared with MIRT results. When MIRT linking was being conducted, it was found that ability distribution changed across the two years. After MIRT linking was completed, one more cut-point, at the 50th percentile, was applied to explore the effects of the location of cut-point in ability distribution on the change of proficiency rates.

3.3.4. MIRT linking

3.3.4.1. MIRT linking – FCIP linking on ability dimension

For dimensionality analysis, two years' test data were analyzed using BMIRT (Bayesian Multivariate Item Response Theory) program (Yao, 2003). BMIRT is the only software program currently available which can estimate MIRT parameters for both dichotomous and polytomous data. The program can implement the multidimensional two-parameter partial credit (M-2PPC) model and the multidimensional three-parameter logistic (M-3PL) model concurrently so that it can be used for mixed format test data. The program estimates the item, examinee, and population distribution parameters by implementing Markov chain Monte Carlo (MCMC) methods using the Metropolis-Hastings sampling algorithm. The program implements both exploratory and confirmatory MIRT analyses. In exploratory analyses, every item is assumed to be loaded onto all dimensions. However, one item must be selected to anchor each dimension to remove rotational indeterminacy. Confirmatory analyses work in a similar way as the usual confirmatory factor analyses by fixing the dimension of sensitivity for each item and setting zero loadings for the other dimensions—i.e., simple structure.

To explore test data structure—i.e., dimensional structure—exploratory data analyses were conducted using BMIRT. To get a stationary MCMC chain, several different options of iteration and burn-in, 10,000/5000, 15,000/5000, 20,000/10,000, 25,000/10,000, and 30,000/10,000 options were tried. After examining standard deviations and parameter tracing files, the 30,000/10,000 option was used for all MIRT analyses.

The choice of population ability correlation is arbitrary, but it serves to fix the scale. Different population variance and covariance options were tried by changing the values a little bit in each calibration run, using a small number of iterations (5000/1000). The population correlation was set as .36 by proposing .15, .42, and .15 as values for ability prior covariance, ability proposal variance, and ability proposal covariance respectively. This decision also was made based on the examination of acceptance rate following the recommendation on acceptance rate by Yao, the BMIRT program developer.

After deciding on the calibration options, three different dimensional solutions were explored, from two to four dimensions. In this examination of dimensional structure, not only model fit indices reported by the BMIRT program – such as Akaike's (1987) information criterion (AIC), BIC, and difference chi-square – were considered, but also loading structures. Model fit indices show that the two-dimensional solution has the best fit in terms of fit indices. However, model fit indices—such as AIC and BIC values—were not considered as the sole criterion for the choice of the dimension. Preliminary cluster analyses of item parameters calibrated from different dimensional solutions were conducted and content of test items clustered together were examined to find out if cluster analyses results produce meaningfully interpretable clusters. Based on this preliminary analysis, the three-dimensional solution was selected for further analyses because the solution produced three relatively sensible item clusters.

After the decision on calibration options for BMIRT was made, five 2006 random sample data sets were calibrated. A conservative burn-in length of 10,000 and a total of 30,000 iterations were employed. Approximate BMIRT running time for the 10,000 (samples) x 36 (items) data set on a Window-based Pentium ®1.78 GHz desktop machine for a single run was about twelve hours. By running the program with an upgraded machine, the run time could be reduced about half. Approximate BMIRT running time for the 10,000 (samples) x 36 (items) data set on a Windows-based IP Intel® Core ™2 CPU T5300@1.78 GHz laptop machine for a single run was about six hours and 10 minutes.

The priors for items parameters were set as follows; μ_a = 1.5 with a normal distribution, and μ_a = 0.0 with a normal distribution, and for *c* parameters *a* = 6 and *b* = 16 for the beta distribution, *beta* (*a, b*) following the solution proposed by Yao and Schwartz (2006).

With the options for BMIRT running as described, first five 2006 RSD sets were calibrated. After getting item parameters from each run, the item parameters and loading structures were compared. Because loading structures and mean values of item parameters across five RSD sets were not consistent, more RSD sets were calibrated to get more stable estimates results. After running 12 RSD sets among 13, five RSD sets showed relatively consistent loading structures and similar item parameter estimates. These five RSD sets were considered to have produced relatively stable estimates and thus selected for the linking study.

After getting item parameters from each run, the item parameters estimated from BMIRT were used as fixed item parameters to run the 2005 RSD sets with both operational test items and field test items. For this MIRT FCIP linking, the BMIRTanchor program was used. BMIRTanchor estimates item parameters based on the scale of the fixed item parameters and also estimates ability on the same metric of the fixed item parameters.

To run the BMIRTanchor, four input files need to be prepared. For each run, a control file, an item response data file, item parameter file for all items—items to be estimated plus items with fixed parameters—and an ability scale score (θ scale score) file are needed. BMIRTanchor does not need a separate "NP" file for not-presented items. By giving value "f" for not-presented items, BMIRTanchor recognizes them as not-presented items. The ability score files as input files were obtained by running BMIRT. To prepare input item parameter (IP) files to run BMIRTanchor, the result of RSD501A was used. It was expected that item parameters of 2005 operational test items would be estimated regardless of the values provided in an input item parameter file, as in PARSCALE. So, the same values of 42 items for the 2005 operation test items were used with changed values of common item parameters from each 2006 RSD set calibration.

Examination of item parameter estimates of 2005 RSD sets obtained by running BMIRTanchor suggested strong effects of the values of 42 items given in the input IP files. When the same estimates for the items to be estimated (i.e., the 2005 operational test items) were used in the input IP files, three anchor items had exactly the same estimates. Based on this finding, all 2005 RSD sets were calibrated to get their own estimates of 42 operational test items. Four more input item parameter files were prepared by running BMIRT. Using new IP files, four more MIRT FCIP linking calibrations were conducted by running BMIRTanchor.

A simulation study to check the precision of BMIRT estimation was conducted. While PARSCALE has been widely used in testing industry and psychometric research community, BMIRT is a relatively new program. There are some research results which report the reliability of BMIRT estimation through simulation studies (Yao & Schwartz, 2006; Yao & Boughton, 2007), but it has not been reported if it also has been used in practical applications. The simulation study was conducted on 28 multiple-choice items from the 2006 test because there was no program available to generate simulated data for

MIRT polytomous response data.[21] First, response data for 28 multiple-choice items by 10,000 random samples from the 2006 test data were calibrated on BMIRT with the same prior options which were used in this study. Using item parameters calibrated by BMIRT as true parameters, 10,000-sample simulated response data were generated from a multivariate normal distribution with mean vector and variance/covariance matrix of the ability estimation data from BMIRT. The simulated data were calibrated using the same options on BMIRT and the item parameter estimates from simulated data were compared with the true item parameters to check item recovery ability of BMIRT.

3.3.4.2. Dimensionality analyses/cluster analyses

After the 2005 test data and the 2006 test data are linked through the FCIP linking method, hierarchical cluster analyses (HCA) was used to explore the test structure. Cluster analysis is a statistical method for segmenting or grouping a collection of objects such as observations, individuals, cases, etc. into relatively homogeneous subsets or clusters so that those within each cluster are more similar to each other than to objects grouped in different clusters, based on the characteristics under consideration. Hierarchical cluster analysis employs more than a single step of partition to form clusters. While there are several algorithms which are commonly used for cluster analysis, Ward's (1963) method was used for cluster analyses for this study. Ward proposed a clustering procedure so that the information loss associated with each grouping is minimized. Information loss is defined by Ward in terms of an error sum of squares criterion.

The hierarchical cluster analyses (HCA) in MIRT use the angular distance between the directions of the best measurement of items, which is converted from

[21] There might be programs for generating simulated data of MIRT polytomous items being used by individual researchers, but they were not available to the researcher and developing the program for the purpose was beyond the goal of this study.

estimated multidimensional discrimination (MDISC) parameter estimates. Application of cluster analysis to MIRT dimensional structure was first proposed by Miller and Hirsch (1992). They used angular distance converted from MDISC as a measure of proximity for clustering. The angles between the directions of item vectors were used as a similarity measure. Kim (2001) showed that Ward's method with the angular distance yielded stable classifications under various test conditions including mixed test format for MIRT cluster analyses.

Hierarchical cluster analysis for each random sample data set, a total of 10 RSD sets, from five cluster analyses for each year, was conducted using a MATLAB program, MIRTCLUST[22]. MIRTCLUST provides a cluster diagram for a set of items (n = 36) in a d-dimensional space (d = 3) based on the angle between the items. The five cluster diagrams for the 2006 five RSD sets are presented below (Figure 3.1 to Figure 3.5).

[22] MIRTCLUST was written by Reckase in 2001.

Figure 3.1 Cluster diagram of RSD 601

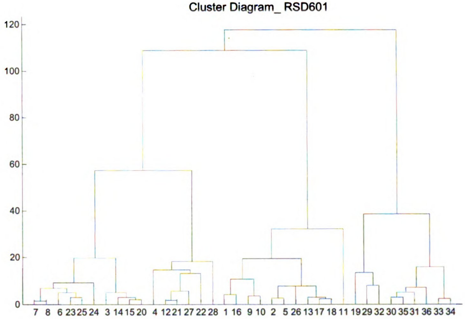


Figure 3.2 Cluster diagram of RSD 602

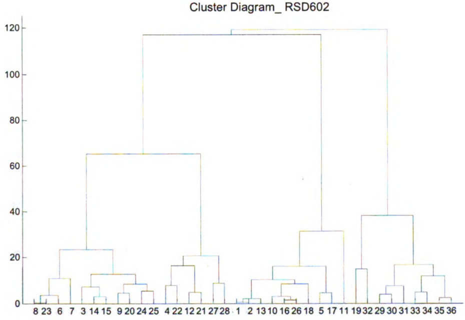


Figure 3.3 Cluster diagram of RSD 603

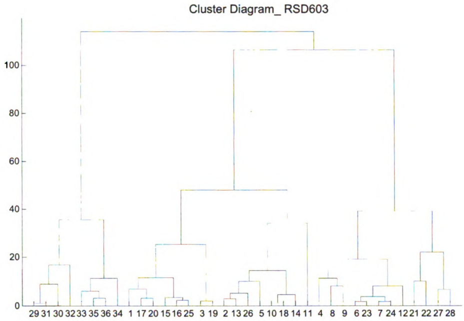


Figure 3.4 Cluster diagram of RSD 604

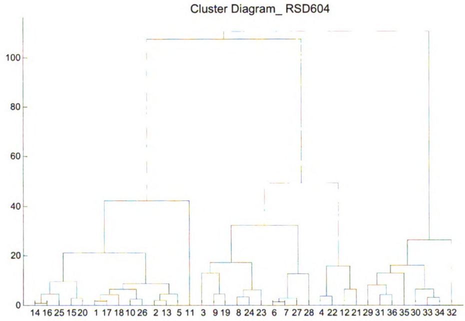
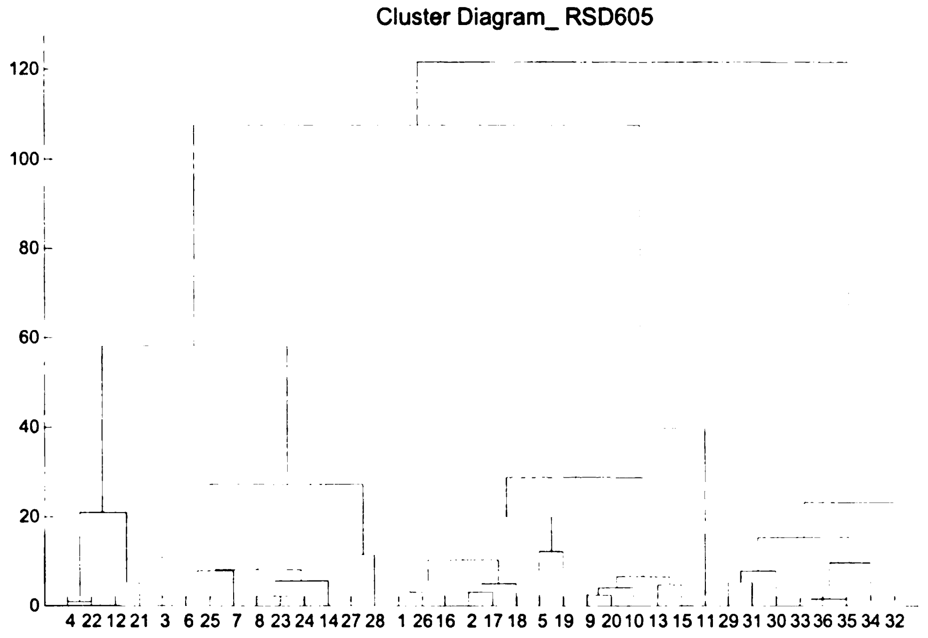


Figure 3.5 Cluster diagram of RSD 605



As shown in the cluster diagrams, three relatively distinct clusters were identified for all five 2006 RSD sets. The clustering was similar across the five RSD sets. For example, eight constructed response (CR) items, item 8 to item 11 and item 27 to item 30, were clustered into one cluster in all five RSD sets, even though there was one additional item, 19, in RSD601 and RSD602. From close examination of cluster diagrams and the values of discrimination parameters, it was found that items in one cluster—the left side cluster in the case of RSD605—have similarly high loadings on θ_1 and θ_2. The second group of items clustered together had high loading on θ_1, and the third group of items clustered together was CR items and had high loadings on θ_3. The three groups of items were notated as cluster 1, 2, or 3 (7th column in Table 3.3). When an item was classified into the same cluster in the all five RSD sets, as the case of item 1 into cluster

Table 3.3 Cluster number of each item for RSDs and for the 2006

No	601	602	603	604	605	2006
1	2	2	2	2	2	2
2	2	2	2	2	2	2
3	1	1	2	1	1	1
4	1	1	1	1	1	1
5	2	2	2	2	2	2
6	1	1	1	1	1	1
7	1	1	1	1	1	1
8	3	3	3	3	3	3
9	3	3	3	3	3	3
10	3	3	3	3	3	3
11	3	3	3	3	3	3
12	1	1	1	1	1	1
13	2	1	1	1	2	1
14	2	2	2	2	2	2
15	2	2	2	2	2	2
16	1	1	1	1	1	1
17	2	2	2	2	2	2
18	1	1	2	2	1	2
19	1	1	2	2	2	2
20	2	2	2	2	2	2
21	2	2	2	2	2	2
22	2	2	2	2	2	2
23	3	3	2	1	2	2
24	1	1	2	2	2	2
25	1	1	1	1	1	1
26	1	1	1	1	1	1
27	3	3	3	3	3	3
28	3	3	3	3	3	3
29	3	3	3	3	3	3
30	3	3	3	3	3	3
31	1	1	1	1	1	1
32	1	1	1	1	1	1
33	1	1	2	2	1	1
34	2	2	2	2	2	2
35	1	1	1	1	1	1
36	1	1	1	1	1	1

2, the rows of the items were shaded and the final clustering group for the 2006 test data was bolded and italicized. When there was a discrepancy across RSD sets, not just cluster diagram but also loading structure was examined. Item 19 presented the biggest problem in making a decision. Cluster analyses did not present a possible solution. Examination of loading structure suggested that the item is close to items in cluster 2.

After the table for the 2006 data was prepared, the same procedure was applied to classify the 2005 test data. Through the cluster analyses using MATLAB, five cluster diagrams were produced and the same table of clustering (Table 3.4) was created in the same way.

Table 3.4 Cluster number of each item for RSDs and for the 2005

Item Type	Item No	501	502	503	504	505	**2005**
MC	1	1	2	2	1	2	2
MC	2	1	1	1	1	1	***1***
MC	3	1	1	1	1	1	***1***
MC	4	3	1	2	2	2	2
MC	5	1	1	1	1	1	***1***
MC	6	1	1	2	1	1	1
MC	7	3	1	3	3	3	3
MC	8	2	1	2	2	2	2
SA	9	1	1	3	1	3	1
SA	10	1	3	3	1	1	1
CR	11	3	3	3	3	3	***3***
CR	12	3	3	3	3	3	***3***
CR	13	3	3	3	3	3	***3***
MC	14	2	2	2	2	2	***2***
MC	15	1	1	3	1	1	1
MC	16	2	2	2	2	2	***2***
MC	17	3	3	3	3	3	***3***
MC	18	2	2	3	2	2	***2***
MC	19	2	1	3	2	2	2
MC	20	1	1	1	1	1	***1***
MC	21	2	2	2	2	2	***2***
MC	22	1	1	1	1	2	1

Table 3.4 (cont'd)

MC	23	2	2	2	2	2	***2***
MC	24	2	2	2	2	2	***2***
MC	25	1	3	3	1	3	1
MC	26	1	1	1	1	1	***1***
MC	27	1	1	1	1	1	***1***
MC	28	2	2	2	2	2	***2***
MC	29	1	2	1	1	2	1
CR	30	3	3	3	3	3	***3***
CR	31	3	3	3	3	3	***3***
SA	32	3	3	3	3	3	***3***
SA	33	3	1	3	2	2	3
MC	34	1	1	1	1	1	***1***
MC	35	1	2	1	1	2	1
MC	36	1	1	1	1	1	***1***
MC	37	1	1	3	1	1	1
MC	38	2	2	2	2	2	***2***
MC	39	1	2	1	1	1	1
MC	40	1	1	1	1	1	***1***
MC	41	2	1	3	2	2	2
MC	42	2	1	3	1	2	2

As the cluster diagrams, Figures 3.6 to 3.10, show, items were clustered into three relatively distinct groups for all five 2005 RSD sets. Depending on whether it is possible to identify meaningful constructs corresponding to clusters, the number of clusters can be selected differently. Overall it seems that the 2005 test data allow not only three-cluster solution but also four-cluster solution, which might present measurement of more specific constructs. In this study, selection of the number of clusters was made with a substantial consideration of the number of constructs identified through the cluster analysis of the 2006 test data. As the cluster diagrams suggest, there were more discrepancies of cluster classifications across the 2005 RSD sets. This was expected because the 2005 RSD set was linked with the 2006 RSD set from 2006 test framework by fixing the scale as the 2006 scale. The final decision was made in the similar way to the 2006 data as described.

Figure 3.6 Cluster diagram for fixed RSD501

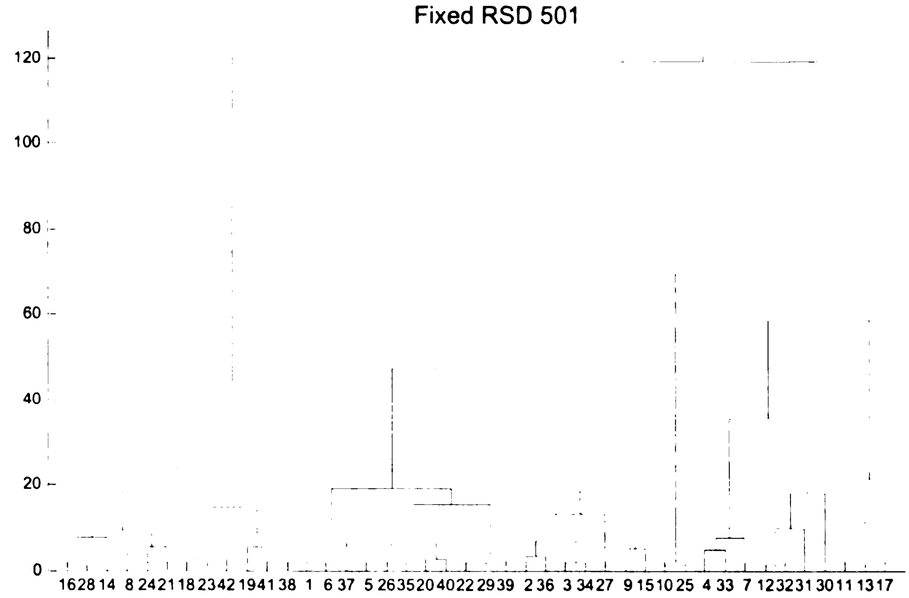


Figure 3.7 Cluster diagram for fixed RSD502

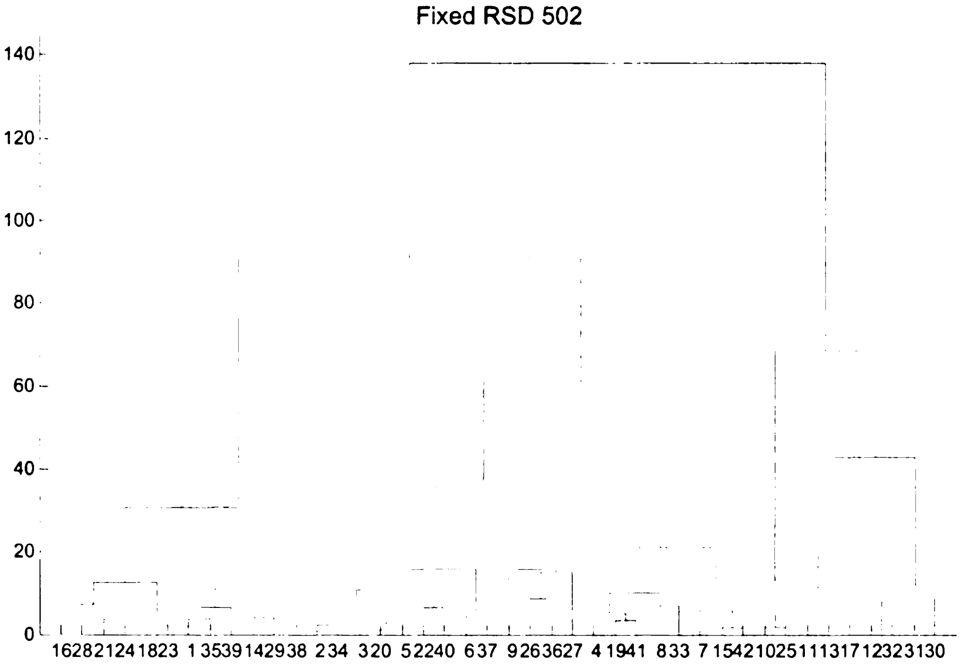


Figure 3.8 Cluster diagram for fixed RSD503

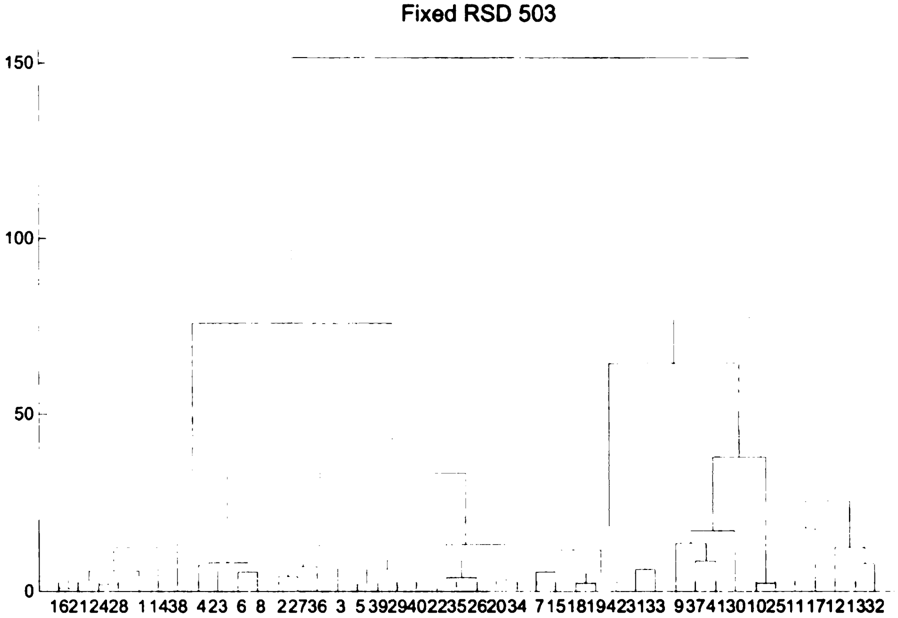


Figure 3.9 Cluster diagram for fixed RSD504

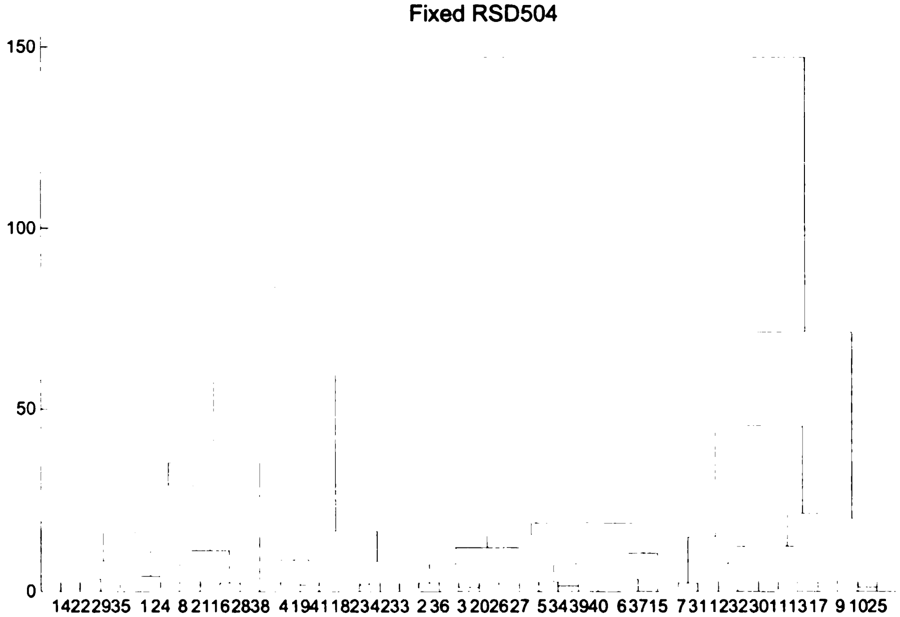
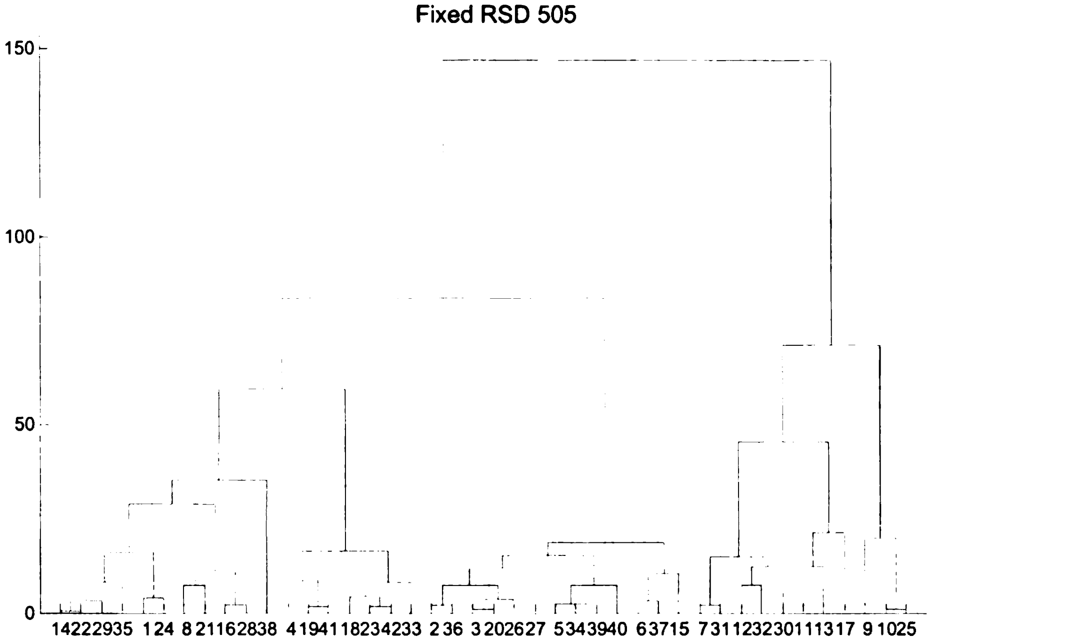


Figure 3.10 Cluster diagram for fixed RSD505



Unlike the 2006 test, there were two multiple-choice items (7 and 17) clustered with cluster 3 items, which had high values on θ_3—mostly constructed-response or short-answer items. Discussion of the relationship between θ dimensions, mathematical ability constructs, which were identified as a result of the item review committee activity, and some sample items, are provided in the next section.

3.3.4.3. Item review committee

Once cluster analyses were complete, the item review committee was convened. The item review committee consisted of four doctoral students with a mathematics education major. The committee members were recruited from the mathematics education learning community at Michigan State University. All of the committee members had experience teaching and working with upper elementary to middle school students[23]. The committee meeting was held in mid August, 2007 at Michigan State University and continued for five and half hours in total.

Following the researcher's introduction to the study and cluster analysis results, cluster diagrams and the tables of clusters along with real test items and test specifications were provided to the committee. Each member reviewed items in terms of content and process standards. It was requested that each member consider what mathematical ability might be employed for students to answer the item correctly from the students' point of view.

After completing a review of individual items—a total 78 items (36 from the 2006 test and 42 from the 2005 test)—two groups of two members each discussed constructs measured by the items assigned to the same cluster. First, naming of the clusters formed by items from the 2006 test was attempted. Next, the groups considered whether these names could be applied to the 2005 test clustering. Each group made its own decision on the meaning of the clusters identified through MIRT cluster analyses. After within-group discussion, results from the two groups were compared and discussed to reach a consensus of the meanings of θ dimensions and the clusters identified. Three identified clusters were judged to be meaningfully interpretable and three mathematical constructs being measured in common by the items within each cluster were determined based on the committee discussion. Brief descriptions of the constructs identified were prepared by the researcher based on the committee discussion. The procedures and results of the item review committee discussion and decision of the constructs and their meanings are summarized as below.

[23] Kosze Lee, Joo-Young Park, Violeta Yurita, and Marcy Wood at Michigan State University served on the Item Review Committee.

First, items grouped as cluster *1* had high *a*-parameters on θ_1, and θ_1 was identified as computational ability and/or knowledge of simple concepts. Most of the items in cluster *1* could be considered to require number sense and number operations. Items in cluster *2* had high *a*-parameters on θ_1 and θ_2, and θ_2 was identified as mathematical thinking or reasoning ability. The items in cluster 2 were judged to require both computational ability and certain level of mathematical reasoning. These items required more than one step of solution and setting up a strategy to solve the problem. These items were different from typical routine problem solving questions in that they do not allow direct application of routine problem solving methods and therefore need setting up a problem solving strategy to solve the items. The items in cluster 3 had higher *a*-parameters on θ_3 compared to θ_1 and θ_2. θ_3 was identified as communication and/or representation abilities. The items in cluster 3 were judged to require communication and/or representation abilities. Table 3.5 presents a brief summary description of each θ dimension.

Table 3.5 Description of θ dimensions

θ_1	Computation, number sense, simple concept, straightforward
θ_2	Mathematical thinking or reasoning; slightly complicated concepts
θ_3	Communication; representation

Second, the mathematical construct measured through the items in cluster 1was named as "problem solving"—a combination of θ_1 and θ_2 as described in Table 3.5. Item 31 in the 2006 test is a typical problem solving item which requires both computational ability and ability to think mathematically to set up a strategy of solution[24]. Item 31 is presented below as an example item measuring problem solving construct.

---

31. In a hockey arena, the first row has 276 seats, the second row has 288 seats and the third row has 300 seats. Each row after this continues to increase by the same number. If the arena has a total of 6 rows, how many seats are in the arena?

a 1836
b 1176
c 972
d 312

To answer this question correctly, a problem-solving strategy needs to be set up because several steps are required to achieve a solution—the question cannot be solved by the application of simple procedural knowledge. Students need to (1) compute the number of seats by which each row increases by subtraction; (2) compute the number of seats for rows 4, 5, and 6 respectively by addition; and (3) add the number of seats of all six rows to get the total number of the seats in the arena. Setting up a problem-solving strategy requires more than computations. It needs mathematical thinking or reasoning to determine which computational procedures are to be applied and when to apply them. This item requires relatively high abilities on both θ_1 and θ_2 and relatively low ability on θ_3.

The mathematical construct measured by the items in cluster 2 was named as "procedural knowledge". An example item in this construct category is presented below.

[24] The three items presented in this dissertation were represented with the permission of EQAO. All question items for the 2005 test and the 2006 test are provided as Appendix attachments.

1. Which is the most appropriate unit of measurement to describe the area of the floor a gym?

a. km^2
b. cm^2
c. m^2
d. m^3

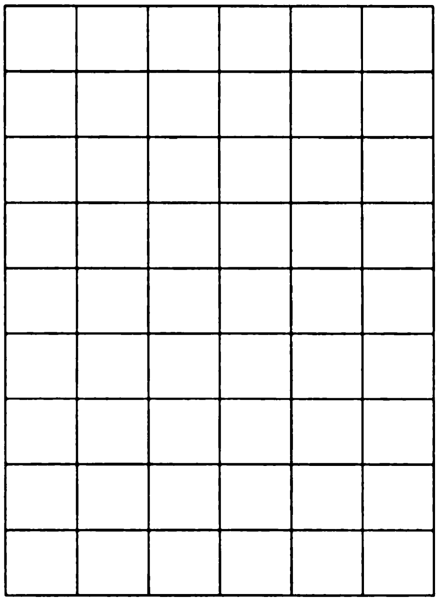
Item 1 in 2006 test presented above asks for the most appropriate unit of measurement to describe the area of the floor of a gym. While this item has been classified as measurement content area item by test specification, it does need simple knowledge and understanding of a simple mathematical concept, e.g. a measurement unit. It does need high ability on θ_1 and lower ability on both θ_2 and $\theta_{3..}$.

Third, the mathematical construct measured by items in cluster 3 was named as communication and representation. One example item for this construct is presented below. Item 27 requires high ability on $\theta_{3.}$and less ability on θ_2, and lowest ability on θ_1. The item requires ability to represent given data for an appropriate communication of the information given in the data.

27. Ranjit makes the chart below to record the amount of money collected during a fundraising event.

Day	Monday	Tuesday	Wednesday	Thursday	Friday
Amount of Money Collected	$50	$125	$75	$25	$175

Make a broken-line graph to represent the data. Remember to include all titles and labels.



Explain your choice of scale.

The classification of items based on the three mathematical constructs for the 2006 test is presented in Table 3.6.

Table 3.6 Assignment of items to constructs: 2006

Item No	601	605	606	607	612	2006
1	PK	PK	PK	PK	PK	**PK**
2	PK	PK	PK	PK	PK	**PK**
3	PS	PS	PK	PS	PS	**PS**
4	PS	PS	PS	PS	PS	**PS**
5	PK	PK	PK	PK	PK	**PK**

Table 3.6 (cont'd)

6	PS	PS	PS	PS	PS	**PS**
7	PS	PS	PS	PS	PS	**PS**
8	CR	CR	CR	CR	CR	**CR**
9	CR	CR	CR	CR	CR	**CR**
10	CR	CR	CR	CR	CR	**CR**
11	CR	CR	CR	CR	CR	**CR**
12	PS	PS	PS	PS	PS	**PS**
13	PK	PS	PS	PS	PK	**PS**
14	PK	PK	PK	PK	PK	**PK**
15	PK	PK	PK	PK	PK	**PK**
16	PS	PS	PS	PS	PS	**PS**
17	PK	PK	PK	PK	PK	**PK**
18	PS	PS	PK	PK	PS	**PK**
19	PS	PS	PK	PK	PK	**PK**
20	PK	PK	PK	PK	PK	**PK**
21	PK	PK	PK	PK	PK	**PK**
22	PK	PK	PK	PK	PK	**PK**
23	CR	CR	PK	PS	PK	**PK**
24	PS	PS	PK	PK	PK	**PK**
25	PS	PS	PS	PS	PS	**PS**
26	PS	PS	PS	PS	PS	**PS**
27	CR	CR	CR	CR	CR	**CR**
28	CR	CR	CR	CR	CR	**CR**
29	CR	CR	CR	CR	CR	**CR**
30	CR	CR	CR	CR	CR	**CR**
31	PS	PS	PS	PS	PS	**PS**
32	PS	PS	PS	PS	PS	**PS**
33	PS	PS	PK	PK	PS	**PS**
34	PK	PK	PK	PK	PK	**PK**
35	PS	PS	PS	PS	PS	**PS**
36	PS	PS	PS	PS	PS	**PS**

MC refers to multiple choice questions.

SA refers to short answer question.

CR refers to constructed response questions.

FCIP linking was conducted with the 2006 test as a reference test, which means that ability space for the 2005 test data was specified as the same ability space as the 2006 test data. Whether this was a feasible approach needs to be examined. First, the decision was made based on a statistical basis. As already explained, the data collection structure made estimation of the anchor items from the 2005 test less stable compared to the opposite direction. But, there is another practical implication of this approach. If tests are expected to measure current students' learning in accordance with the current curriculum framework and instructional approach, which are not expected to remain unchanged, comparison of students' learning across years from the current framework would be more meaningful. The classification of items based on mathematical constructs for the 2005 test is presented in Table 3.7.

Table 3.7 Assignment of items to constructs: 2005

Item Type	Item No	501	502	503	504	505	**2005**
MC	1	PS	PK	PK	PS	PK	**PK**
MC	2	PS	PS	PS	PS	PS	**PS**
MC	3	PS	PS	PS	PS	PS	**PS**
MC	4	CR	PS	PK	PK	PK	**PK**
MC	5	PS	PS	PS	PS	PS	**PS**
MC	6	PS	PS	PK	PS	PS	**PS**
MC	7	CR	PS	CR	CR	CR	**CR**
MC	8	PK	PS	PK	PK	PK	**PK**
SA	**9**	**PS**	**PS**	**CR**	**PS**	**CR**	**PS**
SA	10	PS	CR	CR	PS	PS	**PS**
CR	11	CR	CR	CR	CR	CR	**CR**
CR	12	CR	CR	CR	CR	CR	**CR**
CR	13	CR	CR	CR	CR	CR	**CR**
MC	14	PK	PK	PK	PK	PK	**PK**
MC	15	PS	PS	CR	PS	PS	**PS**
MC	16	PK	PK	PK	PK	PK	**PK**
MC	17	CR	CR	CR	CR	CR	**CR**
MC	18	PK	PK	CR	PK	PK	**PK**

Table 3.7 (cont'd)

MC	19	PK	PS	CR	PK	PK	**PK**
MC	20	PS	PS	PS	PS	PS	**PS**
MC	21	PK	PK	PK	PK	PK	**PK**
MC	22	PS	PS	PS	**PS**	PK	**PS**
MC	23	PK	PK	PK	PK	PK	**PK**
MC	24	PK	PK	PK	PK	PK	**PK**
MC	25	PS	CR	CR	PS	**CR**	**CR**
MC	26	PS	PS	PS	PS	PS	**PS**
MC	27	PS	PS	PS	PS	PS	**PS**
MC	28	PK	PK	PK	PK	PK	**PK**
MC	29	PS	PK	PS	**PS**	PK	**PK**
CR	30	CR	CR	CR	CR	CR	**CR**
CR	31	CR	CR	CR	CR	CR	**CR**
SA	32	CR	CR	CR	CR	CR	**CR**
SA	33	CR	PS	CR	**PK**	**PK**	**PK**
MC	34	PS	PS	PS	PS	PS	**PS**
MC	35	PS	PK	PS	PS	PK	**PS**
MC	36	PS	PS	PS	PS	PS	**PS**
MC	37	PS	PS	CR	PS	PS	**PS**
MC	38	PK	PK	PK	PK	PK	**PK**
MC	39	PS	PK	PS	PS	PS	**PS**
MC	40	PS	PS	PS	PS	PS	**PS**
MC	41	PK	PS	CR	PK	PK	**PK**
MC	42	PK	PS	CR	**PS**	**PK**	**PK**

MC refers to multiple choice questions.

SA refers to short answer question.

CR refers to constructed-response questions.

The constructs identified above can be interpreted psychometrically as the best measurement of the items as a composite of abilities and/or skills required for correct responses within the ability space defined as three ability dimensions. The direction of the best measurement of a cluster of items is called a reference composite, as already explained. The UIRT approach presents only one reference composite for a whole test. The MIRT approach allows finer distinctions, dividing a broadly defined domain such as mathematical ability into several constructs—for example, problem solving, procedural knowledge, and communication and representation, as in the case of this study. How the direction of each reference composite was determined are discussed in the next section.

3.3.4.4. Determining reference composites

Table 3.8 reports the discrimination parameter estimates for three ability dimensions for each item for one 2006 RSD set. For the purpose of illustration, items were rearranged so that items in the same construct can be presented together. PK items have the highest discrimination parameter estimates on θ_2, and PS items usually high on both θ_1 and θ_2 and CR items high on θ_3.

For the purpose of explanation, the **a**-matrix of the construct PK consisting of the first 14 items will be denoted as RC (reference composite) 1, that for PS as RC2, and for CR as RC3. It was shown that the unidimensional item response theory scale is defined by the first eigenvector of the **a'a** matrix which corresponds to the largest eigenvalues of the matrix (Wang, 1985), where **a** is the matrix of discrimination parameters for the compensatory MIRT model.

Table 3.8 Discriminations and RC classification of a 2006 RSD set

	a1	a2	a3	**RC**
1	**0.922**	0.331	0.320	**PK**
2	**1.017**	0.265	0.279	**PK**
5	**1.055**	0.284	0.555	**PK**
14	**0.776**	0.526	0.299	**PK**
15	**2.038**	0.000	0.000	**PK**
17	**1.048**	0.651	0.302	**PK**
18	**1.136**	0.909	0.425	**PK**
19	**0.840**	0.432	0.247	**PK**
20	**0.895**	0.327	0.230	**PK**
21	**1.425**	0.403	0.466	**PK**
22	**1.220**	0.321	0.268	**PK**
23	**0.867**	0.473	0.480	**PK**

Table 3.8 (cont'd)

24	**0.716**	0.408	0.251	**PK**
34	**1.353**	0.567	0.462	**PK**
3	**0.522**	**0.488**	0.343	**PS**
4	**0.266**	**0.546**	0.228	**PS**
6	**1.016**	**0.868**	0.259	**PS**
7	**0.880**	**0.938**	0.246	**PS**
12	**0.871**	**0.783**	0.321	**PS**
13	**0.637**	**0.389**	0.253	**PS**
16	**0.410**	**1.238**	0.227	**PS**
25	0.280	**1.207**	0.214	**PS**
26	0.226	**0.473**	0.205	**PS**
31	**1.121**	**1.041**	0.388	**PS**
32	**0.804**	**0.779**	0.281	**PS**
33	**1.642**	**1.368**	0.385	**PS**
35	**0.916**	**1.421**	0.219	**PS**
36	**1.111**	**1.251**	0.000	**PS**
8	0.268	0.202	**0.433**	**CR**
9	0.575	0.208	**0.602**	**CR**
10	0.334	0.374	**0.691**	**CR**
11	0.270	0.387	**0.519**	**CR**
27	0.339	0.211	**0.449**	**CR**
28	0.217	0.201	**0.314**	**CR**
29	0.278	0.287	**0.548**	**CR**
30	0.329	0.359	**0.644**	**CR**

To find a rotation matrix for RC1, **RC1'RC1** was calculated first. **RC1'RC1** produced the following matrix.

$$\begin{bmatrix} 18.2575 & 6.0855 & 4.8171 \\ 6.0855 & 3.0567 & 2.1231 \\ 4.8171 & 2.1231 & 1.7575 \end{bmatrix}$$

Each diagonal value in the matrix is the sum of the squared *a*-elements for each ability dimension of the RC1-matrix. The off diagonal values are the sums of the cross-products of the *a*-elements from different dimensions. The eigenvalues for this matrix are

0.1727, 1.1283, and 21.7706, which sum is 23.0717. This sum of the eigenvalues is the same as the sum of the diagonal elements. The eigenvector that corresponds to the largest eigenvalue is $\begin{bmatrix} .9110 \\ .3250 \\ .2538 \end{bmatrix}$

As explained in Reckase (in press), the sum of the squared elements of the eigenvector is equal to 1, so the elements have the properties of direction cosines. These direction cosines give the orientation of the reference composite with the coordinate axes of the ability space specified in this study.

The angle between the reference composite and the coordinate axes can be determined by taking the arccosine of the elements of the eigenvector (Reckase, ibid). In this example, the reference composite has an angle of approximately 24° with the θ_1 axis, 71° with the θ_2 axis and 75° with the θ_3 axis, suggesting the reference composite is most strongly related to dimension 1. Applying the same procedure to RC2 and RC3 produces an angle of approximately 50° with the θ_1 axis, 42° with the θ_2 axis and 79° with the θ_3 axis for RC2 suggesting the reference composite is strongly related to both dimension 2 and dimension 3; and an angle of approximately 61° with the θ_1 axis, 66° with the θ_2 axis and 39° with the θ_3 axis for RC3 suggesting the reference composite is most strongly related to dimension 3. Once the directions of reference composites are determined in this way, the coordinate system can be rotated so that the direction of reference composites are aligned with the axis 1 of the rotated system and ability on a reference composite can be measured as a distance along the axis 1 in the new coordinate system. This procedure is explained in the next section.

3.3.4.5 Projecting Ability at Reference Composites as Mathematical Constructs

In the MIRT approach, it is assumed that each item requires more than one ability or proficiency to correctly answer the item. MIRT analysis typically identifies more than one ability dimension, i.e., ability coordinate axis. Then, an ability space being measured is specified by these statistically identified ability coordinate axes. If these statistically identified coordinate axes correspond to meaningful constructs under interest, abilities measured by these coordinate axes will indicate abilities on constructs of interest. However, ability coordinates do not necessarily correspond to meaningful constructs, as indicated by Reckase (2006). For this reason, MIRT cluster analysis was conducted to identify meaningful constructs as in this study. Through hierarchical cluster analyses and careful examination of test contents by content experts, researchers have identified meaningfully interpretable constructs (Martineau et al., 2006; Li, 2006).

If statistically identified ability coordinate axes can be considered as meaningfully interpretable constructs, it would be more useful to measure ability on constructs than on statistically identified coordinate axes. If ability on constructs can be measured, learning growth also can be traced on the constructs rather than on the coordinate axes.

In the case of Li's study (2006), the ability coordinate axes were assumed to correspond to the constructs identified through MIRT hierarchical cluster analysis and abilities measured at the coordinate axes were used as construct abilities. When a discrimination loading structure is close to a simple structure, meaning that only one dimension has high loadings, with close to zero loadings at other dimensions, treating ability coordinates as corresponding to constructs would be acceptable. However, when the loading structure is not close to a simple structure, treating an ability coordinate axis as if it is a construct would result in more measurement error than necessary in measuring constructs. If it is possible to measure ability at the construct level even when ability coordinates do not correspond to constructs identified through MIRT cluster analyses, learning growth at the construct level can be observed and linking also could be conducted at the construct ability level.

Recently, Reckase (in press) showed a way of measuring ability in the direction given by a vector of direction cosines of an item, which is the best measurement as a composite of multiple abilities and/or skills of the item in a multiple ability space. For example, if an item has the best measurement in the direction of 30 degrees from coordinate axis 1 in a two-dimensional case, the direction of the best measurement of the item can be aligned with a new coordinate axis 1 by rotating the current coordinate axis 1 by 30 degrees counter-clockwise. Then, the new coordinate axis 1 is aligned with the direction of the best measurement of the item. In this way, distances along the rotated coordinate axis 1 now become a direct measure of the best measurement of the item. When ability space is specified by more than two coordinate axes, matching up the direction of the best measurement of an item is achieved through multiple rotations, the number of dimension minus 1.

The same procedure developed by Reckase can be applied to reference composites so that coordinate axes are aligned with reference composites in a new coordinate system in the ability space specified. Reckase also (in press) shows that the invariant property of MIRT model holds after multiple rotations with an example of one item. The specific procedures for rotating reference composites to align the reference composites with the coordinate axes and thus identify ability at reference composite (i.e., construct abilities) are explained below.

Conversion of coordinates in an ability space to a different set of rotated coordinate axes is done by multiplication of the initial coordinates by a rotation matrix. For the two-dimensional case, the rotation matrix is given by

$$\mathbf{Rot} = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix},$$

where α is the number of degrees of rotation in the clockwise direction. To get the rotation matrix in the counter-clockwise direction, negative values of α are used.

Determining angles of rotation in a higher dimensional space than two-dimensional case is much more challenging. Reckase (in press) shows that it is useful to separate a particular rotation into a series of rotations around each of the orthogonal coordinate axes and then to get the full rotation as the product of each of the separate rotation matrices. In general, the angle of rotation needed in each of the θ_l, θ_v-planes is given by

$$\gamma_{1v} = \arccos\left[\frac{\sqrt{\sum_{i=1}^{v-1}\cos^2\alpha_i}}{\sqrt{\sum_{i=1}^{v}\cos^2\alpha_i}}\right] = \arccos\left[\frac{\sqrt{\sum_{i=1}^{v-1}a_i^2}}{\sqrt{\sum_{i=1}^{v}a_i^2}}\right], \text{ for } v = 2, \ldots, m.$$

Both item and person parameters after rotation are obtained by post-multiplying the **a** and **θ** matrices by the rotation matrix, **a Rot** and **θ Rot**.

The rotation matrix as a product of each of the separate rotation matrices for the above example, 2006 RSD, in three-dimensional space under study is given by

$$\begin{bmatrix} 0.9110 & -0.3360 & -0.2390 \\ 0.3250 & 0.9418 & -0.0853 \\ 0.2538 & 0 & 0.9673 \end{bmatrix}$$

The results of the application of the rotation to the **a**-matrix are given in Table 3.9. There are some important results about the a-parameters after rotation. First, each item in cluster 1—the items shaded—has predominantly high loadings on a_1. This is a result of rotating the θ_1-axis to align with the direction of RC1. When an item has the very similar direction to RC1, it has close to a zero discrimination parameter estimate for the other ability dimensions. While most of items are on a direction very similar to RC1, some items with a non-zero discrimination parameter estimate on the other two dimensions are not.

A second result is that the first a-parameter for each item is now very similar to the multidimensional discrimination for the item, which is presented in the 6th column (A), because the discrimination along RC1 is the most discriminating direction. When an item has close to zero discrimination parameter estimates on a_2 and a_3, the value of a_1 is very close to A as in the case of item 1or 20 (bolded and italicized).

Table 3.9 Discriminations and MDISC (A) after rotation

Item	RC	a_1	a_2	a_3	*A*
1	1	**1.03**	0.00	0.06	***1.03***
2	1	**1.08**	-0.09	0.00	1.09
5	1	**1.19**	-0.09	0.26	1.23
14	1	**0.95**	0.24	0.06	0.98
15	1	**1.86**	-0.68	-0.49	2.04
17	1	**1.24**	0.26	-0.01	1.27
18	1	**1.44**	0.47	0.06	1.51
19	1	**0.97**	0.12	0.00	0.98

Table 3.9 (cont'd)

20	1	**0.98**	0.01	-0.02	***0.98***
21	1	**1.55**	-0.10	0.08	1.55
22	1	**1.28**	-0.11	-0.06	1.29
23	1	**1.07**	0.15	0.22	1.10
24	1	**0.85**	0.14	0.04	0.86
34	1	**1.53**	0.08	0.08	1.54
3	2	0.72	0.28	0.17	0.79
4	2	0.48	0.43	0.11	0.65
6	2	1.27	0.48	-0.07	1.36
7	2	1.17	0.59	-0.05	1.31
12	2	1.13	0.44	0.04	1.21
13	2	0.77	0.15	0.06	0.79
16	2	0.83	1.03	0.02	1.32
25	2	0.70	1.04	0.04	1.26
26	2	0.41	0.37	0.10	0.56
31	2	1.46	0.60	0.02	1.58
32	2	1.06	0.46	0.01	1.15
33	2	2.04	0.74	-0.14	2.17
35	2	1.35	1.03	-0.13	1.71
36	2	1.42	0.81	-0.37	1.67
8	3	0.49	0.08	0.34	0.60
9	3	0.34	0.12	0.23	0.43
10	3	0.42	0.10	0.34	0.55
11	3	0.74	0.00	0.43	0.86
27	3	0.60	0.24	0.56	0.85
28	3	0.50	0.27	0.40	0.70
29	3	0.49	0.18	0.44	0.68
30	3	0.58	0.23	0.51	0.81

The same rotation matrix was used to rotate the ability matrix onto the RC1 direction so that ability after rotation could be expressed as ability on RC1 by post-multiplying the $\boldsymbol{\theta}$ matrix.

As an illustration, ability scores of 10 students on original ability dimensions and ability scores after rotating to RC1 are presented in Table 3.10. The values of the first column of the $\boldsymbol{\theta}$ matrix after rotation (in bold) are the ability scores for the first reference composite.

Table 3.10 Coordinates of person locations before and after rotation

	Before rotation			After rotation		
	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
1	0.466	0.316	-0.514	**0.3968**	0.141	-0.6355
2	-0.413	-1.202	-0.86	**-0.9852**	-0.9933	-0.6306
3	-1.504	-0.708	-1.017	**-1.8584**	-0.1614	-0.5639
4	-1.19	-1.856	-1.308	**-2.0193**	-1.3482	-0.8225
5	0.811	0.307	0.552	**0.9787**	0.0166	0.3139
6	0.292	0.858	-0.192	**0.4962**	0.71	-0.3287
7	0.755	0.702	0.982	**1.1652**	0.4075	0.7095
8	0.957	0.281	-0.144	**0.9266**	-0.0569	-0.392
9	-1.642	-0.721	-1.518	**-2.1155**	-0.1273	-1.0144
10	-0.098	0.26	0.484	**0.1181**	0.2778	0.4694

Because three reference composites were identified, three rotation matrices were made, applying the same procedures explained above. The original ability matrix was transformed three times using the three rotation matrices. After these procedures, the fully rotated ability matrix, which is aligned with reference composites, was made by taking values of the first column for each rotated ability matrix. A part of the fully rotated ability matrix is presented in Table 3.11 with an example of 10 students.

Table 3.11 Fully rotated ability matrix

	θ_1	θ_2	θ_3
1	0.3968	0.4381	-0.0458
2	-0.9852	-1.3184	-1.3566
3	-1.8584	-1.6853	-1.8014
4	-2.0193	-2.3881	-2.3448
5	0.9787	0.854	0.9437
6	0.4962	0.788	0.3417
7	1.1652	1.1913	1.4116
8	0.9266	0.7982	0.4632
9	-2.1155	-1.8779	-2.2618
10	0.1181	0.2203	0.4346

Once the full rotated matrices for all five 2006 RSD sets were made, the same procedures were applied to the five 2005A RSD sets to get rotated ability scores so that change of proficiency rate based on RC abilities between 2005 and 2006 could be calculated. One adjustment was, however, made for the 2005 test data. Because two items (items 10 and 25) had unreasonably high discrimination values of about 5, they were considered as outliers and deleted from the analyses. By deleting two items, the direction of three RCs of the 2005 data became more similar to that of 2006, suggesting that reference composites from two years' test data are compatible.

3.3.4.6 MIRT approach to proficiency classification—compensatory vs. conjunctive

In the case of MIRT, there are three achievement areas, i.e., three constructs, to be considered. Therefore, two approaches or models to proficiency classification can be considered— compensatory and conjunctive. Earlier discussions on compensatory versus conjunctive model to decision making can be found in Mehrens and Phillips (1989) in relation to teacher licensure or in Jager (1991) in relation to school quality evaluation. More recently, the two models have been discussed in the context of NCLB accountability system (Paulsen et al., 2002; Hill & DePascale, 2002; Abedi, 2004). All these studies discuss and clarify the different results that different model selection to NCLB Adequate Yearly Progress (AYP) decision making would bring about. In proficiency classification, the compensatory approach assumes that achievement on one construct can compensate for achievement on the other constructs. The conjunctive approach to proficiency classification requires students to achieve "proficiency" in each domain or construct.

The need for the conjunctive approach can be raised by those who are concerned about the possibility that students can achieve proficiency even when they are not proficient on some important domains or constructs. For example, if the cut-point for proficiency is set at 80% correct score and there are 40% easy computational items, 40% easy mathematical concept items, and 20% difficult problem solving items, students can pass the cut-point even though they answer less than 50% of problem-solving items correctly if they average 90% correct in the other areas. This is an extreme example, but it illustrates the concern raised.

In a compensatory approach to proficiency classification, there is one additional consideration—how to weight the three abilities. One possible approach is to consider the weight of each construct as corresponding to the relative percent of raw scores for each construct based on test specifications. Another approach is to consider the relative importance of each construct. There were 14 multiple-choice items each for RC1 and RC2 and 8 constructed-response items in RC3 in the case of the 2006 test. Because constructed-response items had a 4 point scale, RC3 had a larger percentage of the raw score than either RC1 or RC2. On the other hand, procedural knowledge (RC1) and problem-solving abilities (RC2) were considered as important mathematical constructs as communication and representation ability (RC3).

Based on this consideration, it was decided to give equal weight to each construct in the compensatory approach. This decision was applied to the 2005 data because FCIP linking was made with the 2006 test as the reference test. For the conjunctive approach, three separate cut-points for each construct were set at both the 80th and 50th percentiles.

Student who passed all of the three cut-points were classified as having achieved proficiency.

3.3.4.7. Proficiency rate change

To compare the proficiency rate across years, the same standard, that is, the same cut-point should be applied, assuming that scores across years are on the same scale. In real situations, the cut-point for a proficiency level is set by a standard-setting process. For the purpose of this study—comparing proficiency rates between 2005 and 2006—, the proficiency rate was set at the 20th percentile in 2005, assuming that 80% of 2005 students achieved proficiency in the case the UIRT and MIRT compensatory approaches to proficiency classification. The ability score, i.e., θ, at the 20th percentile was then considered as the cut-point for proficiency. This ability score was applied to the 2006 data to get the proficiency rate in 2006. In addition to the 80 percent cut-point, the 50th percentile cut point was also applied, to explore the effects of cut-points on proficiency classification, as already described. In the conjunctive approach, the same percentile criteria were applied to each construct.

CHAPTER 4

RESULTS AND INTERPRETATION

This chapter describes the results of the study and provides the plausible interpretations of the analysis results.

4.1 UIRT Linking

To evaluate educational performance across years using test scores, it is necessary to link test scores from different tests so that results from different tests are put on a common metric and thus can be compared. Unidimensional linking approaches assume that tests are measuring the same unidimensional construct or a composite of common constructs which can be considered unidimensional. As already explained, the procedure for linking employed in this study was very similar to the fixed common item parameter equating design. Other than assuming that two years' tests are different forms of the same test with difference only in difficulty, the basic linking procedure is the same as the equating procedures described in the theoretical framework chapter.

4.1.1. Descriptive statistics of data

Descriptive statistics from the basic item analysis result and estimated unidimensional item parameters for the whole response data set are reported in Table 4.1 (2005) and Table 4.2 (2006). Calibration of the data using PARSCALE was conducted using the same options which were used by EQAO.[25] When the 2005 test data were calibrated, Item 8 had extreme difficulty and standard error, -15.77 and 37.353

[25] The calibration options were as follows; CALIB PARTIAL, LOGISTIC, CYCLE= (500, 1, 1, 1, 1), ITEMFIT=20, NEWTON=30, CRIT=0.0001, NQPT=40.

respectively. The 2005 data were recalibrated again with the increased iterations to 1,000. The increase in the number of iterations decreased the difficulty and standard error by about half. The data were recalibrated again with the iterations set at 3,000, but the difficulty and standard error stayed almost the same as in the calibration using 1,000 iterations. Because of its extreme value, the item 8 was dropped from the analyses. The IRT item parameter estimates reported in Table 4.1 are from the calibration with 1,000 iterations. Parameter estimates for the other items were very similar to those from the original calibration. IRT item parameter estimates and the classical test theory item indices of the 2006 test are reported in Table 4.2.

Table 4.1 Item analysis result—item discrimination, item difficulty, and guessing parameters in item response theory and classical test theory (2005 test, N=132,021)

IRT Item Parameters -- PARSCALE Results							Item parameters: classical test theory	
Item number	Slope	S.E.	Location	S.E.	Guessing	S.E.	P-Value	Biserial
1	0.96	0.014	-0.92	0.022	0.24	0.01	0.77	0.47
2	0.86	0.011	-0.83	0.02	0.13	0.01	0.71	0.45
3	0.79	0.014	0.57	0.013	0.21	0.005	0.47	0.47
4	0.59	0.018	0.48	0.039	0.46	0.009	0.66	0.38
5	0.91	0.009	-0.42	0.011	0.04	0.005	0.59	0.27
6	0.79	0.013	-0.04	0.02	0.28	0.007	0.62	0.53
7	0.67	0.018	1.45	0.017	0.23	0.004	0.38	0.4
8	Deleted from the calibration							
9	0.73	0.005	-0.41	0.006	0	0	0.56	0.26
10	1	0.007	-0.21	0.005	0	0	0.53	0.5
11	0.36	0.002	-0.9	0.005	0	0	2.29	0.58
12	0.32	0.002	-1.76	0.007	0	0	2.82	0.63
13	0.29	0.001	-0.82	0.006	0	0	1.97	0.57
14	0.8	0.013	0.29	0.014	0.19	0.005	0.51	0.55
15	0.87	0.015	0.83	0.011	0.18	0.004	0.4	0.42
16	0.75	0.01	-1.26	0.028	0.06	0.014	0.76	0.38
17	0.73	0.013	0.51	0.015	0.19	0.005	0.48	0.45
18	0.83	0.014	0.78	0.011	0.15	0.004	0.39	0.39
19	0.95	0.018	0.45	0.014	0.35	0.004	0.57	0.4
20	1.01	0.013	-0.32	0.013	0.18	0.006	0.63	0.37

Table 4.1 (cont'd)

21	0.75	0.012	-1.27	0.037	0.17	0.017	0.79	0.5
22	0.79	0.009	-1.63	0.026	0.04	0.015	0.82	0.41
23	0.6	0.01	-0.55	0.031	0.08	0.012	0.61	0.42
24	0.66	0.012	-0.51	0.035	0.27	0.012	0.69	0.42
25	0.91	0.012	-0.03	0.013	0.18	0.005	0.57	0.38
26	0.65	0.012	-0.17	0.028	0.21	0.01	0.61	0.46
27	0.89	0.015	0.78	0.01	0.18	0.004	0.4	0.39
28	0.72	0.012	-0.46	0.028	0.26	0.01	0.68	0.38
29	1	0.015	0.83	0.008	0.13	0.003	0.35	0.4
30	0.35	0.001	-0.66	0.006	0	0	1.81	0.41
31	0.41	0.002	-1.04	0.005	0	0	2.26	0.56
32	0.22	0.004	-1.46	0.026	0	0	0.62	0.65
33	0.79	0.007	-1.8	0.011	0	0	0.84	0.21
34	0.83	0.006	0.61	0.006	0	0	0.32	0.41
35	0.75	0.013	0.44	0.015	0.21	0.005	0.5	0.5
36	1.01	0.014	0.54	0.009	0.17	0.003	0.44	0.38
37	0.84	0.012	0.8	0.009	0.1	0.003	0.35	0.43
38	1.1	0.015	-0.77	0.017	0.29	0.008	0.77	0.42
39	1.1	0.012	-0.19	0.01	0.12	0.004	0.57	0.54
40	0.91	0.012	-0.91	0.02	0.15	0.01	0.74	0.46
41	0.61	0.013	0.23	0.027	0.21	0.009	0.54	0.36
42	1.01	0.015	0.26	0.012	0.26	0.004	0.55	0.44

Table 4.2 Item analysis result -- item discrimination, item difficulty, and guessing parameters in item response theory and classical test theory (2006 test, N=136,653)[26]

IRT item parameters -- PARSCALE results							Item parameters in classical test theory	
Item	Slope	S.E.	Location	S.E.	Guessing	S.E.	P-Value	Bi-serial
1	0.77	0.012	-0.895	0.037	0.243	0.016	0.812	0.35
2	0.74	0.010	0.124	0.017	0.096	0.007	0.562	0.29
3	0.64	0.011	0.015	0.030	0.185	0.011	0.621	0.45
4	0.44	0.011	-0.111	0.064	0.140	0.018	0.615	0.25
5	0.94	0.019	1.441	0.011	0.204	0.003	0.353	0.34
6	1.03	0.013	-0.966	0.024	0.232	0.013	0.850	0.28
7	1.03	0.012	-0.912	0.020	0.151	0.012	0.821	0.37
8	0.47	0.002	-1.827	0.006	0	0	0.775	*0.52*
9	0.31	0.002	-1.538	0.007	0	0	0.681	*0.43*
10	0.40	0.002	-1.537	0.006	0	0	0.731	*0.52*

[26] In some items, guessing and standard errors of guessing parameter estimates are zero because they are constructed response items.

Table 4.2 (Cont'd)

11	0.62	0.002	-0.398	0.005	0	0	0.537	*0.57*
12	0.92	0.01	-0.447	0.015	0.074	0.008	0.697	0.36
13	0.62	0.01	-0.197	0.033	0.156	0.012	0.652	0.27
14	0.84	0.017	0.845	0.016	0.335	0.005	0.555	0.41
15	1.12	0.016	1.09	0.008	0.150	0.003	0.355	0.47
16	0.81	0.008	-1.064	0.021	0.003	0.011	0.791	0.49
17	0.95	0.018	0.888	0.013	0.310	0.004	0.524	0.38
18	1.13	0.016	-0.128	0.015	0.329	0.007	0.730	0.21
19	0.71	0.010	-0.065	0.020	0.080	0.008	0.599	0.43
20	0.74	0.012	-0.626	0.036	0.267	0.014	0.776	0.41
21	1.09	0.012	0.001	0.011	0.128	0.005	0.617	0.33
22	0.80	0.015	1.343	0.011	0.147	0.004	0.342	0.44
23	0.85	0.012	0.008	0.017	0.169	0.007	0.629	0.28
24	0.69	0.010	-0.142	0.024	0.122	0.010	0.634	0.42
25	0.75	0.010	-0.924	0.033	0.124	0.016	0.791	0.41
26	0.17	0.022	0	1.265	0	0.148	0.829	0.40
27	0.59	0.002	-0.755	0.006	0	0	0.658	*0.52*
28	0.53	0.003	-1.113	0.005	0	0	0.738	*0.60*
29	0.50	0.002	-0.871	0.005	0	0	0.643	*0.58*
30	0.57	0.003	-0.441	0.005	0	0	0.582	*0.58*
31	1.23	0.016	0.300	0.01	0.223	0.004	0.591	0.35
32	0.92	0.013	-0.010	0.016	0.215	0.007	0.656	0.41
33	1.57	0.015	0.455	0.006	0.075	0.002	0.457	0.37
34	1.19	0.016	1.024	0.007	0.141	0.003	0.360	0.41
35	1.10	0.013	-0.522	0.016	0.190	0.008	0.766	0.56
36	1.16	0.015	-0.865	0.02	0.247	0.011	0.854	0.46

Table 4.3 and Table 4.4 present the summary statistics for item parameters of the 2005 test and the 2006 test. Summary statistics after the 2005 test data were calibrated with fixed parameters for the anchor items are presented in Table 4.5. When the 2005 test was linked with the 2006 test by fixing the item parameter estimates from the 2006 test calibration (Table 4.5), average item discrimination (slope) was very similar between the two tests, but the average item difficulty (threshold) increased by .15, suggesting the 2005 test was more difficult than the 2006 test.

Table 4.3 Summary statistics for parameter estimates (2005)

PARAMETER	MEAN	SD	N
SLOPE	0.76	0.22	41
THRESHOLD	-0.23	0.82	41
GUESSING	0.14	0.11	31

Table 4.4 Summary statistics for parameter estimates (2006)

PARAMETER	MEAN	SD	N
SLOPE	0.80	0.29	36
THRESHOLD	-0.25	0.81	36
GUESSING	0.18	0.08	27

Table 4.5 Summary statistics for parameter estimates (2005A)

PARAMETER	MEAN	SD	N
SLOPE	0.79	0.24	36
THRESHOLD	0.08	0.78	36
GUESSING	0.14	0.11	27

Table 4.6 and Table 4.7 present item parameter estimates by the item type. In both the 2005 test and the 2006 test, item discrimination of multiple-choice items is higher than that of constructed-response items. Short-answer items in the 2005 test have a little lower discrimination on average.

Table 4.6 Item parameter estimates by item type for the 2005 test*

	Mean			Standard deviation		
	MC	CR	SA	MC	CR	SA
Discrimination	0.86	0.36	0.75	0.16	0.05	0.31
Difficulty	0.11	-0.89	-0.48	0.71	0.41	0.93
Guessing	0.19	0.00	0.00	0.08	0.00	0.00

*MC refers to multiple choice items, CR to constructed-response items, and SA to short-answer items.

Table 4.7 Item parameter estimates by item type for the 2006 test*

	Mean		Standard deviation	
	MC	CR	MC	CR
discrimination	0.89	0.50	0.27	0.10
difficulty	-0.01	-1.06	0.72	0.53
guessing	0.17	0.00	0.09	0.00

*MC refers to multiple choice items and CR to constructed-response items.

The information functions provided by PARSCALE suggest that polytomous items provide more information overall. This finding confirms the results of previous research, which suggests higher information in constructed-response items than multiple-choice items (Donohue, 1993).

4.1.2. Proficiency rate change between year 2005 and year 2006

While the five pairs of random sample data sets with sample size of 10,000 were linked in this study for the comparison of UIRT linking and MIRT linking, the whole data sets were also linked through the FCIP method and change of proficiency rates was explored for the UIRT linking to check the randomness of the RSD sets, i.e., representativeness of the samples in each RSD set[27].

Mean, standard deviation, maximum and minimum θ values for five RSD sets and all students in 2005 are reported in Table 4.8. The reported statistics for the 2005 test are based on the data linked to the 2006 data so that the statistics can be directly compared to those for the 2006 data. θ values in 2005 are from 0.015 to .186 lower than in 2006; .134 lower on average for RSD sets and .139 lower for all students. However, there was no statistically significant difference between 2005 RSD sets and their matched 2006 RSD

[27] Because each RSD set was separately calibrated, they cannot be compared to each other directly. Representativeness of samples mean how the samples represent (i.e., resemble) the population. The similarity of descriptive statistics of RSD set with the parameters suggests a piece of evidence for the representativeness of the samples.

sets at α=0.01. Overall mean and standard deviations across five RSD sets are very close to those for the all students in 2005. Even though there are some variations across the RSD sets in each year, there was no statistically significant difference across the RSD sets in both years. One interesting finding here is that ranges of θ values in 2005 for both the students in RSD sets and all students have been shrunken especially for students with low θ values. This shrinkage might be the result of estimation procedure rather than changes in the ability distribution. This phenomenon is more salient for MIRT linking case, which will be discussed in the MIRT linking result report section.

Table 4.8 Mean, standard deviation, maximum, and minimum of θ for 2005 test data

	501A	*502A*	*503A*	*504A*	*505A*	*Mean_RSD*	*2005A_All*
Mean	-0.186	-0.167	-0.162	-0.140	-0.015	-0.134	-0.139
SD	1.020	1.003	1.002	0.990	0.959	0.995	0.990
Max	2.677	2.645	2.653	2.749	2.730	2.691	2.761
Min	-3.369	-3.262	-3.338	-3.225	-3.246	-3.288	-3.350

As a way of checking the effects of distribution change, two different proficiency cut-points were explored; the 20th percentile, assuming that 80% of 2005 students reached the proficiency level, and the 50th percentile, assuming that 50% of 2005 students reached the proficiency level. Proficiency rate changes are reported in Table 4.9 for each paired RSD sets and for all students in 2005 and 2006.

Table 4.9 Cut scores for proficiency level and proficiency rates in 2006

	501-601	*502-602*	*503-603*	*504-604*	*505-605*	*Mean_RSD*	*All students*
The 20th percentile	-0.8506	-0.8516	-0.8189	-0.8417	-0.8487	-0.8407	-0.8136
passing in 2006	*79.93%*	*79.82%*	*79.41%*	*80.15%*	*80.17%*	*79.90%*	*79.37%*
The 50th percentile	*0.0038*	*0.0109*	*0.0264*	*0.0148*	*0.0119*	0.01356	*0.0326*
passing in 2006	*51.53%*	*51.13%*	*50.70%*	*51.22%*	*50.98%*	*51.11%*	*50.54%*

When the proficiency cut point was set at the 20th percentile for 2005 students, the proficiency rates in 2006 were 79.90% on average for the five 2005 RSD sets and 79.37% for the whole student data. This result suggests that the proficiency rate from 2005 to 2006 changed little when the proficiency level was set at the 20th percentile of the 2005students. The results of χ^2 test confirmed that there was no significant difference between 2005 and 2006 across the five RSD sets (Table 4.10).

Table 4.10 χ^2 results_UIRT_20th Percentile

UIRT_20th percentile

	χ^2
501-601	0.02
502-602	0.10
503-603	1.08
504-604	0.07
505-605	0.09

When the proficiency level was set at the 50th percentile, the result was similar (Table 4.9). While RSD sets showed about 1% increase in the proficiency rate in 2006, there was only .5% increase in the whole students' data. The results of χ^2 test suggest that only one RSD pair (501-601) among the five has statistically significant difference in proficiency rate between 2005 and 2006. Overall, when students' performances from two years were linked through the unidimensional linking approach, the result suggests little improvement of performance in 2006 compared to 2005 and the cut-point did not make difference in the change of proficiency rates.

Table 4.11 χ^2 results_UIRT_50th Percentile

UIRT_50th percentile	
	χ^2
501-601	4.68
502-602	2.55
503-603	0.98
504-604	2.98
505-605	1.92

In this study, FCIP linking was conducted from 2006 to 2005, the opposite of the usual direction, because of the test data structure with missing data for the 2005A test. This means that students' performance in 2005 was measured relative to the 2006 calibration. Students who were classified as proficient based on the 2005 test result could fail to achieve the proficiency level when their proficiency was classified from the 2006 test scale. This can happen because estimates of item parameters from each separate calibration might be a little different from those resulted from the FCIP linking, which in turn can make a little different ability estimates for individual students. To check this effect, the number and percentage of misclassifications for all students at the 20th percentile cut-point was explored. Table 4.12 presents the proficiency rates from both directions—from the 2005 test perspective and the 2006 test perspective. The result suggests that inconsistency in proficiency classification as a group is almost the same when linking in either direction but individual students might be advantaged or disadvantaged depending on the direction in which linking occurs. Because the previous year's proficiency rate is determined prior to the next year, this result provides a justification for adopting a reverse direction FCIP linking method.

Table 4.12 Proficiency classification by FCIP linking from both directions

Passing	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No-Pass on either	25460	19.28	25460	19.28
No pass on 2005 and pass on 2005A	941	0.71	26401	20
No pass on 2005A and pass on 2005	942	0.71	27343	20.71
Pass on both	104678	79.29	132021	100

4.2 MIRT Linking

4.2.1. Checking BMIRT estimation precision—a simulation study

While there is some research documenting the stability of BMIRT estimation (Yao & Schwartz, 2005; Yao & Schwartz, 2006), the program is relatively new (developed in 2003) and there is no research using the program other than by the program developer and colleagues. To check the item parameter recovery ability of BMIRT, one simulation study was conducted. The simulation was conducted for dichotomous data—multiple-choice items—only because polytomous item-response-data generation in MIRT framework was not available.

First, the response data for the 28 dichotomous items out of the 36 items in the 2006 test with 10,000 samples were calibrated by BMIRT. The estimates of the item parameters calibrated were treated as "true" parameters. Using MATLAB, 10,000 simulated response strings (10,000 X 28) were generated from multivariate normal distribution with the mean vector and variance/covariance matrix from the calibration. Then, the simulated data were calibrated by BMIRT with the same prior options which were used to calibrate the data. The result showed that BMIRT recovered item parameters reasonably well. The "true" item parameters and the estimated item parameters from the simulated data are reported in Table 4.13.

Table 4.13 True item parameters and recovered item parameters

	True_	Rec_	True_	Rec_	True_	Rec_	True_	Rec_	True_	Rec_
item	a1	a1	a2	a2	a3	a3	b	b	c	c
1	0.81	0.7	0.25	0.36	0.51	0.57	-1.46	-1.5	0.22	0.23
2	0.97	1.13	0.24	0.29	0.42	0.53	-0.1	-0.24	0.07	0.12
3	0.54	0.49	0.28	0.38	0.44	0.53	-0.3	-0.33	0.12	0.14
4	0.27	0.23	0.48	0.48	0.32	0.38	-0.27	-0.35	0.15	0.17
5	1.12	1.01	0.25	0.3	0.53	0.67	2.07	1.97	0.2	0.25
6	0.78	0.84	0.68	0.73	0.79	0.62	-2.27	-2.26	0.12	0.12
7	0.72	0.82	0.85	1.1	0.92	1.02	-2.09	-2.38	0.11	0.12
8	0.73	0.77	0.67	0.86	0.67	0.85	-1.05	-1.34	0.07	0.1
9	0.6	0.56	0.31	0.42	0.39	0.38	-0.48	-0.71	0.15	0.2
10	0.9	0.79	0.36	0.38	0.27	0.28	0.66	0.74	0.29	0.24
11	2.23	2.35	0	0	0	0	1.84	1.95	0.13	0.1
12	0.31	0.31	1	1.13	0.85	0.83	-1.84	-1.97	0.06	0.07
13	1.25	1.08	0.35	0.32	0.63	0.94	1.44	1.67	0.34	0.27
14	1.02	0.99	0.74	0.59	0.39	0.65	-0.8	-0.83	0.25	0.26
15	0.72	0.51	0.37	0.45	0.49	0.67	-0.35	-0.39	0.08	0.1
16	0.87	0.61	0.28	0.3	0.37	0.75	-1.11	-1.08	0.23	0.2
17	1.43	1.43	0.39	0.54	0.6	0.75	-0.45	-0.68	0.11	0.14
18	1.01	1.21	0.25	0.29	0.45	0.45	1.65	1.49	0.15	0.27
19	0.89	0.86	0.37	0.45	0.39	0.44	-0.32	-0.44	0.15	0.18
20	0.71	0.63	0.44	0.55	0.37	0.5	-0.44	-0.58	0.11	0.14
21	0.43	0.36	0.94	1.11	0.35	0.34	-1.51	-1.57	0.11	0.11
22	0.22	0.28	0.57	0.55	0.22	0.29	-1.58	-1.77	0.16	0.18
23	1.18	1.17	1.07	1.13	0.25	0.4	0.07	-0.11	0.2	0.23
24	1.05	1.09	0.75	0.83	0.25	0.35	-0.5	-0.65	0.19	0.23
25	1.5	1.6	1.27	1.32	0.4	0.48	0.65	0.56	0.07	0.1
26	1.39	1.35	0.45	0.53	0.39	0.43	1.55	1.43	0.12	0.17
27	0.92	0.8	1.44	1.31	0.33	0.41	-1.52	-1.33	0.18	0.16
28	1.05	1.03	1.52	1.63	0	0	-2.53	-2.45	0.21	0.2
Mean	0.92	0.89	0.59	0.66	0.43	0.52	-0.39	-0.47	0.16	0.17
SD	0.428	0.461	0.391	0.402	0.217	0.249	1.286	1.31	0.069	0.06

Root mean squared error (RMSE) was calculated for each item and the means of RMSE for three discriminations, difficulty, and guessing parameters are reported in Table 4.14. Table 4.15 reports the correlations between the true and the recovered item parameters. Compared to the amount of RMSE usually reported in simulation studies[28], the RMSE values were evaluated as relatively small, which confirmed the item recovery ability of BMIRT. The correlations between true item parameters and recovered ones by BMIRT range from .99(*b*) to .86(*c*) (Table 4.14). Plots of a_1(r =.97) and a_3 (r =.89) are presented in Figures 4.1 and 4.2 to graphically indicate the similarities between true values and recovered ones.

Table 4.14 RMSE of item discriminations

a1	a2	a3	b	c
0.0850	0.0880	0.1052	0.1278	0.0306

Table 4.15 Correlations between the true and the recovered item parameters

		a1_true	a2_true	a3_true	b_true	c_true
a1_simu	Pearson Correlation	**0.97**				
		0.00				
a2_simu	Pearson Correlation		**0.98**			
	N					
a3_simu	Pearson Correlation			**0.89**		
b_simu	Pearson Correlation				**0.99**	
c_true	Pearson Correlation					**0.86**
**	Correlation is significant at the 0.01 level (2-tailed).					

[28] For example, Ball et al. (2002) report about .14 to .20 of RMSE of each item for the 30 items test with 500 samples. The much smaller values of RMSE in this study seem to be due to the large sample size.

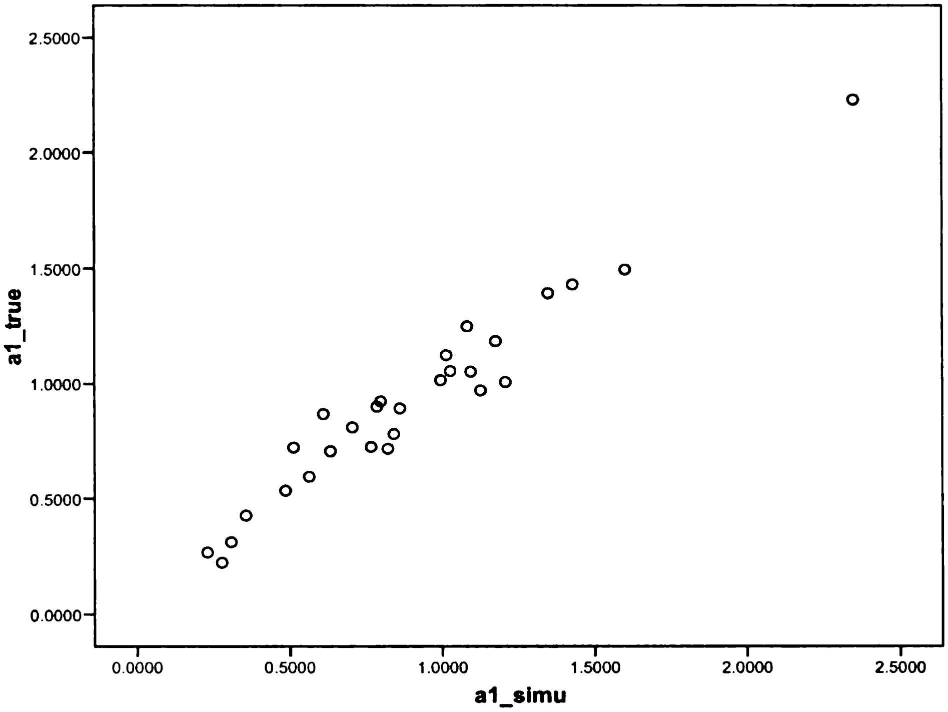


Figure 4.1 Scatter plot of a recovered discrimination, a_1, and true values

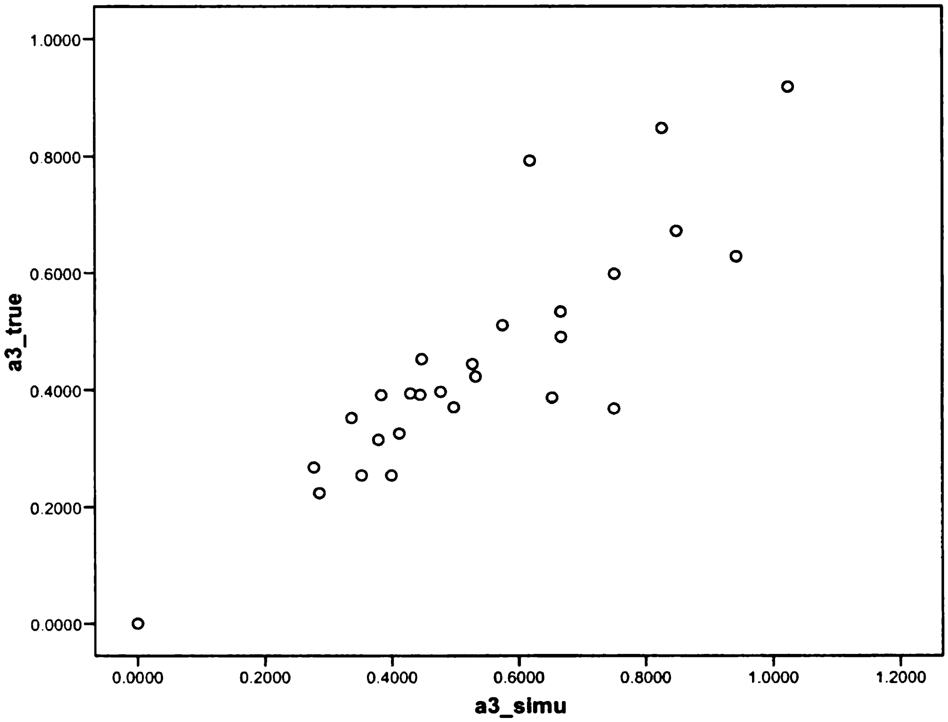


Figure 4.2 Scatter plot of a recovered discrimination, a_3, and true values

4.2.2. MIRT calibration results

Five 2006 RSD sets were calibrated by running BMIRT with the same options as described in the Methods chapter. Means of three discrimination values for each RSD set and means and standard deviations of discriminations across the five RSD sets are presented in Table 4.16. Overall discriminations across the five RSD sets for each dimension are relatively similar.

Table 4.16 Mean of item discriminations for the 2006 RSD sets

	a1	a2	a3
RSD601	0.84	0.58	0.34
RSD602	0.85	0.60	0.35
RSD603	0.78	0.64	0.33
RSD604	0.82	0.64	0.32
RSD605	0.79	0.58	0.34
Mean	0.82	0.61	0.34
SD	0.03	0.03	0.01

Table 4.17 reports MDISC for 36 items for the five RSD sets. Discriminations range from 0.42 for Item 30 to 2.80 for Item 11. Overall, discriminations of constructed-response items are low compared to those of the multiple-choice items, suggesting either high or low MDIFF, which is analogous to difficulty in UIRT. When items are too difficult or too easy, they do not discriminate well overall. Table 4.18 reports MDIFF for the 36 items for the five 2006 RSD sets. Average MDIFF of the multiple-choice items (-0.33) is much higher than that of the constructed-response items (-1.35). The difference in MDIFF and MDSIC between MC items and CR items suggests the possibility that they measure different constructs, which will be discussed in the cluster analysis report section.

Table 4.17 MDSIC for 36 items in 2006 Test

Item	Item Type	RSD601	RSD602	RSD603	RSD604	RSD605
1	MC	1.04	1.17	1.01	1.07	1.03
2	MC	1.13	1.05	1.03	1.04	1.09
3	MC	0.81	0.84	0.74	0.93	0.79
4	MC	0.64	0.69	0.57	0.68	0.65
5	MC	1.35	1.45	1.17	1.21	1.23
6	MC	1.40	1.41	1.30	1.37	1.36
7	MC	1.42	1.45	1.50	1.34	1.31
8	CR	*0.62*	*0.61*	*0.61*	*0.62*	*0.60*
9	CR	*0.44*	*0.42*	*0.42*	*0.42*	*0.43*
10	CR	*0.55*	*0.57*	*0.54*	*0.57*	*0.55*
11	CR	*0.90*	*0.90*	*0.90*	*0.86*	*0.86*
12	MC	1.24	1.27	1.26	1.27	1.21
13	MC	0.80	0.73	0.83	0.79	0.79
14	MC	1.05	1.09	1.08	1.20	0.98
15	MC	2.27	2.51	2.39	2.80	2.04
16	MC	1.27	1.33	1.20	1.28	1.32
17	MC	1.39	1.28	1.36	1.28	1.27
18	MC	1.36	1.68	1.53	1.66	1.51
19	MC	0.97	0.94	0.95	0.99	0.98
20	MC	1.04	1.12	0.94	1.03	0.98
21	MC	1.61	1.70	1.40	1.47	1.55
22	MC	1.19	1.04	1.02	1.12	1.29
23	MC	1.14	1.03	1.12	1.00	1.10
24	MC	0.92	0.90	0.95	0.88	0.86
25	MC	1.14	1.28	1.19	1.25	1.26
26	MC	0.64	0.59	0.58	0.57	0.56
27	CR	*0.85*	*0.79*	*0.82*	*0.80*	*0.85*
28	CR	*0.77*	*0.75*	*0.77*	*0.72*	*0.70*
29	CR	*0.70*	*0.73*	*0.70*	*0.77*	*0.68*
30	CR	*0.82*	*0.85*	*0.84*	*0.81*	*0.81*
31	MC	1.60	1.85	1.83	1.61	1.58
32	MC	1.35	1.38	1.25	1.38	1.15
33	MC	1.94	2.28	2.04	2.15	2.17
34	MC	1.59	1.68	1.66	1.51	1.54
35	MC	1.78	1.69	1.94	1.79	1.71
36	MC	1.91	1.91	1.84	1.84	1.67
	Mean_all	1.16	1.19	1.15	1.17	1.12
	Mean_MC	1.28	1.33	1.27	1.30	1.25
	Mean_CR	0.71	0.70	0.70	0.70	0.68

Table 4.18 MDIFF for 36 items in 2006 test

Item	Item Type	RSD601	RSD602	RSD603	RSD604	RSD605
1	MC	-1.36	-1.31	-1.64	-1.51	-1.56
2	MC	-0.09	-0.08	-0.13	-0.1	-0.14
3	MC	-0.14	-0.11	-0.33	-0.07	-0.24
4	MC	-0.33	-0.02	-0.38	-0.35	-0.19
5	MC	1.54	1.56	1.55	1.63	1.59
6	MC	-1.53	-1.57	-1.66	-1.62	-1.66
7	MC	-1.42	-1.38	-1.31	-1.55	-1.49
8	CR	-2.58	-2.51	-2.77	-2.53	-2.6
9	CR	-2.14	-2.1	-2.16	-2.09	-2.15
10	CR	-2.13	-2.05	-2.17	-2.11	-2.18
11	CR	-0.66	-0.67	-0.63	-0.68	-0.71
12	MC	-0.82	-0.82	-0.8	-0.8	-0.86
13	MC	-0.61	-0.73	-0.4	-0.57	-0.55
14	MC	0.75	0.65	0.8	0.81	0.65
15	MC	0.83	0.8	0.78	0.86	0.86
16	MC	-1.38	-1.32	-1.46	-1.37	-1.29
17	MC	0.95	0.75	0.79	0.92	0.82
18	MC	-0.52	-0.27	-0.49	-0.25	-0.35
19	MC	-0.3	-0.28	-0.34	-0.3	-0.31
20	MC	-0.88	-0.94	-1.26	-0.99	-1.17
21	MC	-0.27	-0.17	-0.27	-0.27	-0.29
22	MC	1.43	1.57	1.55	1.45	1.37
23	MC	-0.12	-0.28	-0.24	-0.35	-0.21
24	MC	-0.48	-0.4	-0.41	-0.44	-0.54
25	MC	-1.24	-1.07	-1.3	-1.2	-1.24
26	MC	-2.38	-2.17	-2.43	-2.61	-2.68
27	CR	-1.14	-1.16	-1.23	-1.16	-1.18
28	CR	-1.62	-1.58	-1.59	-1.6	-1.68
29	CR	-1.29	-1.24	-1.24	-1.23	-1.34
30	CR	-0.73	-0.71	-0.74	-0.76	-0.76
31	MC	0.09	0.19	0.13	0.15	0.13
32	MC	-0.29	-0.08	-0.24	-0.15	-0.45
33	MC	0.35	0.34	0.33	0.32	0.35
34	MC	1.01	1.04	0.99	1.03	1.07
35	MC	-0.82	-0.89	-0.76	-0.73	-0.92
36	MC	-1.25	-1.28	-1.28	-1.31	-1.46
	Mean_all	-0.60	-0.56	-0.63	-0.60	-0.65
	Mean_MC	-0.33	-0.30	-0.36	-0.33	-0.38
	Mean_CR	-1.35	-1.31	-1.38	-1.33	-1.39

Correlations of MDSIC and MDIFF between the five RSD sets from the 2006 test are in Table 4.19 and Table 4.20 respectively. Correlations of both MDISC and MDIFF are very high. As expected from low MDSIC values for CR items, MDIFF of the items were very high. This supports the random equivalence of the five 2006 RSD sets because both MDISC and MDIFF between equivalent test forms are usually very high and MDIFF are even higher.

Table 4.19 Correlations of MDSIC between RSD sets of the 2006 test

	RSD601	RSD602	RSD603	RSD604	RSD605
RSD601	1				
RSD602	0.978	1			
RSD603	0.9805	0.9748	1		
RSD604	0.967	0.9783	0.9718	1	
RSD605	0.9758	0.9751	0.9677	0.9575	1

Table 4.20 Correlations of MDIFF between RSD sets of the 2006 test

	RSD601	RSD602	RSD603	RSD604	RSD605
RSD601	1				
RSD602	0.9945	1			
RSD603	0.9947	0.9914	1		
RSD604	0.9962	0.9939	0.9943	1	
RSD605	0.9956	0.9936	0.9942	0.9964	1

4.2.3 FCIP MIRT linking on ability dimension

Once item response data of the five 2006 RSD sets were calibrated by the BMIRT program, the five 2005 RSD sets were linked through the FCIP linking method by running the BMIRTanchor program on the data sets. Following the procedures described in the Methods chapter, the five 2005 RSD sets were calibrated and then each 2005 RSD set was linked with its pre-matched 2006 RSD set with the anchor items fixed as the item parameters of from each matched 2006 RSD set.

Item parameter summary statistics for the 42 items for the 2005 RSD sets—RSD501A to RSD505A—after the FCIP linking are reported in Table 4.21. Because the parameter estimates of the 2005 test items and the 2006 test items are on the same metric through the FCIP linking, the values of item discrimination of the 2005 test can be compared directly with those of the 2006 test. The item discriminations of the three dimensions from the 2006 test were 0.82, 0.61, and 0.34 each (Table 4.16). The average item discriminations for the three dimensions in the 2005 RSD sets are overall much higher than those from the 2006 test. Because item discrimination of the Item 10 and Item 25 were unreasonably high—about 6 and 5 each—these two items were deleted from the further analyses.

Table 4.21 Mean of item discriminations of 2005A RSD sets: 42 items

	a_1	a_2	a_3
501A	1.15	1.08	0.69
502A	1.16	1.05	0.69
503A	1.22	1.02	0.72
504A	1.09	1.08	0.60
505A	1.10	0.98	0.74
Mean	1.14	1.04	0.69
SD	0.05	0.04	0.05

Mean of item discriminations for the three dimensions after deleting the two items are reported in Table 4.22. The mean discrimination for a_1 changed little (from 1.14 to 1.17), that of a_2 decreased dramatically (from 1.04 to 0.59), and that of a_3 increased somewhat (from 0.60 to 0.81). Compared to the 2006 test (Table 4.15), a_1 and a_3 are higher and a_2 are similar in the 2005 test. Mean discriminations in the 2006 test were 0.82, 0.61, and 0.34 respectively.

Table 4.22 Mean of item discriminations of 2005A RSD sets: 40 items

	a_1	a_2	a_3
501A	1.17	0.60	0.87
502A	1.18	0.59	0.82
503A	1.24	0.60	0.80
504A	1.12	0.53	0.85
505A	1.13	0.61	0.73
Mean	1.17	0.59	0.81
SD	0.04	0.03	0.05

MDISC for the five 2005A RSD sets after deleting the two items are reported in Table 4.23 below. As expected from high discrimination values for the 2005 test items compared to those of the 2006 test, the average value of MDSIC for the 2005 test items is higher than that of the 2006 test items—1.65 in the 2005 test compared to 1.16 in the 2006 test. Discriminations in the 2005 test are also different by item type as in the 2006 test; highest for MC items (1.78) followed by SA items (1.59), and lowest for CR items (0.92) suggesting lower MDIFF of CR items than two other types of items.

Table 4.23 MDISC for 40 items in the 2005 test

Item	Type	RSD501A	RSD502A	RSD503A	RSD504A	RSD505A
1	MC	1.85	1.82	2.01	1.94	1.91
2	MC	2.55	2.27	2.46	2.13	2.10
3	MC	1.85	1.92	1.92	1.97	1.61
4	MC	1.15	1.02	0.95	1.11	1.09
5	MC	2.14	2.12	2.37	2.28	2.00
6	MC	1.65	1.82	1.62	1.39	1.54
7	MC	1.59	1.59	1.22	1.44	1.45
8	MC	0.91	0.92	0.91	0.84	0.85
9	SA	2.00	2.24	2.34	2.30	2.27
10	***SA***	Deleted				

Table 4.23 (cont'd)

11	CR	1.23	1.17	1.25	1.22	1.40
12	CR	0.74	0.74	0.71	0.72	0.83
13	CR	0.70	0.72	0.77	0.71	0.69
14	MC	1.46	1.54	1.66	1.58	1.61
15	MC	2.23	1.59	1.93	1.84	1.87
16	MC	1.46	1.46	1.51	1.53	1.51
17	MC	1.70	1.52	1.53	1.61	1.78
18	MC	1.58	1.48	1.54	1.62	1.49
19	MC	1.87	1.66	1.76	1.54	1.61
20	MC	2.56	2.33	2.40	2.43	2.28
21	MC	1.52	1.56	1.53	1.45	1.42
22	MC	1.73	1.80	1.81	1.80	1.52
23	MC	1.16	1.17	1.20	1.15	1.11
24	MC	1.28	1.34	1.31	1.21	1.21
25	***MC***	Deleted				
26	MC	1.43	1.59	1.43	1.41	1.51
27	MC	2.80	3.22	2.99	2.84	2.27
28	MC	1.35	1.29	1.41	1.31	1.37
29	MC	2.21	2.19	2.42	2.13	1.97
30	CR	1.00	1.03	0.94	0.90	0.90
31	CR	0.87	0.94	0.87	0.83	0.89
32	SA	0.54	0.54	0.58	0.55	0.50
33	SA	1.51	1.53	1.67	1.42	1.53
34	SA	2.22	2.09	2.11	2.10	1.83
35	MC	1.84	1.65	1.78	1.60	1.50
36	MC	2.79	2.90	2.49	2.55	2.25
37	MC	1.86	1.87	2.20	1.97	1.83
8	MC	2.14	2.17	2.02	2.35	2.09
39	MC	2.58	2.35	2.53	2.27	2.46
40	MC	2.44	2.23	2.18	2.06	2.35
41	MC	1.11	1.11	1.54	1.05	1.23
42	MC	1.82	1.95	2.21	1.74	1.81

The mean MDISC of all items and by item type of each RSD set and across the five 2005 RSD sets are reported in Table 4.24.

Table 4.24 Mean of MDISC by item type across the 2005 RSD sets

Type	RSD501A	RSD502A	RSD503A	RSD504A	RSD505A	Mean	SD
Mean_all	**1.69**	**1.66**	**1.70**	**1.62**	**1.59**	**1.65**	**0.05**
Mean_MC	1.83	1.79	1.83	1.75	1.70	**1.78**	**0.06**
Mean_SA	1.57	1.60	1.68	1.59	1.53	**1.59**	**0.05**
Mean_CR	0.91	0.92	0.91	0.88	0.94	**0.91**	**0.02**

The values of MDIFF of the 40 items after deleting Item 10 and Item 25 in the 2005 test items are overall lower than those of the 2006 test, with the mean of -0.08 compared to -0.61 in the 2006 test items (Table 4.25). CR items have the lowest MDIFF values on average and MC items the highest MDIFF values. Overall, MDIFF values of the 2005 test items are much higher than those of the 2006 test.

Table 4.25 MDIFF for 40 items in the 2005 test

Item	Type	RSD501A	RSD501A	RSD501A	RSD501A	RSD501A	Mean	SD
1	MC	-0.68	-0.75	-0.64	-0.74	-0.68	-0.70	0.05
2	MC	-0.45	-0.38	-0.39	-0.54	-0.54	-0.46	0.08
3	MC	0.49	0.62	0.47	0.45	0.54	0.51	0.07
4	MC	0.43	0.50	0.30	0.46	0.54	0.45	0.09
5	MC	-0.19	-0.14	-0.15	-0.22	-0.24	-0.19	0.04
6	MC	0.05	0.20	0.09	-0.03	0.05	0.07	0.08
7	MC	1.38	1.42	1.28	1.42	1.19	1.34	0.10
8	MC	-1.11	-1.09	-1.12	-1.08	-1.17	-1.11	0.04
9	SA	-0.15	-0.09	-0.07	-0.16	-0.16	-0.13	0.04
11	CR	-0.36	-0.37	-0.30	-0.37	-0.33	-0.35	0.03
12	CR	-1.18	-1.16	-1.19	-1.26	-1.03	-1.16	0.08
13	CR	-0.44	-0.40	-0.33	-0.43	-0.42	-0.40	0.04
14	MC	0.33	0.41	0.40	0.40	0.38	0.38	0.03
15	MC	0.64	0.85	0.80	0.77	0.76	0.76	0.08
16	MC	-0.88	-0.92	-0.83	-0.90	-0.84	-0.87	0.04
17	MC	0.58	0.53	0.64	0.62	0.67	0.61	0.05
18	MC	0.84	0.90	0.88	0.87	0.93	0.88	0.03
19	MC	0.48	0.56	0.54	0.46	0.51	0.51	0.04
20	MC	-0.15	-0.09	-0.15	-0.16	-0.17	-0.14	0.03
21	MC	-0.85	-0.91	-0.83	-0.89	-0.91	-0.88	0.04
22	MC	-1.11	-1.01	-1.00	-1.10	-1.18	-1.08	0.08

Table 4.25 (cont'd)

23	MC	-0.30	-0.25	-0.23	-0.24	-0.28	-0.26	0.03
24	MC	-0.37	-0.17	-0.31	-0.37	-0.30	-0.30	0.08
26	MC	-0.09	0.05	0.01	-0.08	-0.04	-0.03	0.06
27	MC	0.45	0.56	0.56	0.50	0.53	0.52	0.05
28	MC	-0.26	-0.32	-0.24	-0.41	-0.20	-0.29	0.08
29	MC	0.75	0.85	0.76	0.88	0.88	0.82	0.06
30	CR	-0.29	-0.23	-0.26	-0.33	-0.32	-0.29	0.04
31	CR	-0.71	-0.59	-0.67	-0.73	-0.70	-0.68	0.05
32	SA	-0.93	-0.81	-0.76	-0.92	-0.95	-0.87	0.08
33	SA	-1.47	-1.39	-1.30	-1.56	-1.44	-1.43	0.10
34	SA	0.46	0.56	0.52	0.51	0.58	0.53	0.05
35	MC	0.37	0.50	0.53	0.44	0.55	0.48	0.07
36	MC	0.46	0.50	0.46	0.44	0.48	0.47	0.02
37	MC	0.76	0.83	0.75	0.60	0.77	0.74	0.09
38	MC	-0.71	-0.65	-0.68	-0.55	-0.70	-0.66	0.06
39	MC	-0.04	-0.01	0.01	-0.05	-0.01	-0.02	0.02
40	MC	-0.51	-0.53	-0.53	-0.57	-0.52	-0.53	0.02
41	MC	0.28	0.39	0.33	0.41	0.39	0.36	0.05
42	MC	0.40	0.47	0.41	0.35	0.35	0.40	0.05
Mean_all		-0.10	-0.04	-0.06	-0.10	-0.08	-0.08	0.06
Mean_MC		0.03	0.09	0.07	0.04	0.06	0.06	0.06
Mean_SA		-0.52	-0.43	-0.40	-0.53	-0.49	-0.48	0.07
Mean_CR		-0.61	-0.55	-0.55	-0.62	-0.56	-0.58	0.05

The correlations of MDSIC and MDIFF of the 2005 test items across the five RSD sets are presented in Tables 4.26 and 4.27. Again, the correlations of MDSIC and MDIFF are very high.

Table 4.26 Correlations of MDISC between RSD sets of the 2005 test

	RSD501A	RSD502A	RSD503A	RSD504A	RSD505A
RSD501A	1				
RSD502A	0.9856	1			
RSD503A	0.9857	0.9885	1		
RSD504A	0.9865	0.9921	0.99	1	
RSD505A	0.9738	0.9801	0.9812	0.9845	1

Table 4.27 Correlations of MDIFF between RSD sets of the 2005 test

	RSD501A	RSD502A	RSD503A	RSD504A	RSD505A
RSD501A	1				
RSD502A	0.9955	1			
RSD503A	0.9956	0.995	1		
RSD504A	0.995	0.9929	0.9929	1	
RSD505A	0.9943	0.9918	0.9929	0.9925	1

4.2.4 Dimensionality analyses/cluster analyses

The results of cluster analyses suggested that there are three relatively distinctive clusters of items for both tests. These three clusters were identified as three constructs through the item review committee activity as explained in the Methods chapter—C&R (communication and representation), PK (procedural knowledge), and PS (problem solving). The construct measured by each item along with the item type for the 2005 test (40 items) is presented in Table 4.28. The table is arranged so that item type can be easily compared for each construct. The question of whether constructed-response items are measuring different construct than MC items was examined through dimensionality analyses.

In the case of the 2005 test, all the five CR items measure the C&R construct, but two MC items and one SA item are also included in this cluster. Three SA items are distributed across three constructs, while most of the MC items are in either the PK or PS construct, with two items in the C&R construct. Both of these items include graphical representations in the questions. Understanding graphical representation is an important ability necessary to solve the items. As shown in the loading structure, item discriminations, on the right three columns, these items have high loadings on θ_3, as do other items clustered in the R&C construct. This result suggests that item type or format is related to the constructs the items are measuring. This relationship is more salient in the 2006 test (Table 4.27). In the case of the 2006 test, there is no MC item in C&R construct category. However, this does not mean that MC items cannot measure certain constructs, as shown in the example of MC items in the C&R construct category in 2005. An interesting result is that there are no CR items in either the PK or PS construct. This seems to be related to economic considerations rather than the limitation of item type in measuring construct. If PK or PS can be measured as well by MC items, it will be more economical to use MC items than polytomous items. However, there might be the cases that measuring specific knowledge/skills in the PK or PS needs polytomous items.

Table 4.28 Construct and item type for the 2005 test items

Item Type	Item No	**Construct**	Content* strand	Item discrimination		
				a1	a2	a3
CR	11	**C&R**	DMP	0.21	0.31	**1.2**
CR	12	**C&R**	GSS	0.3	0.35	**0.59**
CR	30	**C&R**	M	0.3	0.62	**0.65**
CR	31	**C&R**	NS	0.45	0.47	**0.59**
CR	13	**C&R**	PA	0.23	0.25	**0.63**
MC	7	**C&R**	GSS	**0.88**	0.77	**0.86**
MC	17	**C&R**	DMP	0.65	0	**1.5**
SA	32	**C&R**	GSS	0.21	0.27	**0.41**
MC	21	**PK**	DMP	**1.39**	0.28	0.45
MC	23	**PK**	DMP	**0.97**	0.31	0.53
MC	29	**PK**	DMP	**1.9**	0.98	0.42
MC	1	**PK**	GSS	**1.67**	0.74	0.44
MC	16	**PK**	GSS	**1.36**	0.3	0.5
MC	28	**PK**	GSS	**1.23**	0.29	0.44
MC	38	**PK**	M	**2.15**	0	0
MC	42	**PK**	M	**1.41**	0.73	**1.03**
MC	18	**PK**	NS	**1.24**	0.53	0.72
MC	19	**PK**	NS	**1.34**	0.66	0.76
MC	41	**PK**	NS	**0.87**	0.53	0.61
MC	4	**PK**	PA	**0.8**	0.43	0.54
MC	8	**PK**	PA	**0.76**	0.31	0.32

Table 4.28 (cont'd)

MC	14	**PK**	PA	**1.43**	0.54	0.36
MC	24	**PK**	PA	**1.2**	0.29	0.3
SA	33	**PK**	PA	**1.05**	0.55	**0.96**
MC	5	**PS**	DMP	**1.64**	**1.36**	0.46
MC	34	**PS**	DMP	**1.42**	**1.42**	0.44
MC	40	**PS**	DMP	**1.74**	**1.38**	0.33
MC	26	**PS**	GSS	**1.05**	**0.93**	0.4
MC	2	**PS**	M	**1.5**	**1.68**	0.45
MC	15	**PS**	M	**1.15**	**1.2**	0.87
MC	20	**PS**	M	**1.75**	**1.56**	0.48
MC	3	**PS**	NS	**1.21**	**1.3**	0.51
MC	6	**PS**	NS	**1.26**	**0.83**	0.52
MC	22	**PS**	NS	**1.45**	**0.85**	0.39
MC	27	**PS**	NS	**1.59**	**2.28**	0.41
MC	35	**PS**	NS	**1.36**	**0.87**	0.4
MC	36	**PS**	NS	**1.57**	**1.93**	0.69
MC	37	**PS**	PA	**1.25**	**1.27**	0.75
MC	39	**PS**	PA	**1.99**	**1.33**	0.47
SA	9	**PS**	NS	0.83	**1.79**	*1.02*

*These content strand classification is based on the assessment framework. DMP—Data Management and Probability; GSS—Geometry and Spatial Sense; M—Measurement; NS—Number Sense; PA—Patterning and Algebra

Table 4.29 Construct and item type of the 2006 test items*

Item type	Item No	Construct	Content Strand	Item discrimination		
				a1	a2	a3
CR	29	**C&R**	PA	0.33	0.26	**0.48**
CR	30	**C&R**	GSS	0.21	0.20	**0.33**
CR	31	**C&R**	PA	0.26	0.23	**0.43**
CR	32	**C&R**	M	0.58	0.22	**0.60**
CR	33	**C&R**	M	0.37	0.35	**0.61**
CR	34	**C&R**	PA	0.35	0.33	**0.51**
CR	35	**C&R**	PA	0.32	0.26	**0.54**
CR	36	**C&R**	DMP	0.35	0.34	**0.70**
MC	1	**PK**	M	**0.86**	0.43	0.31
MC	2	**PK**	M	**0.91**	0.44	0.31
MC	5	**PK**	NS	**1.06**	0.45	0.49
MC	10	**PK**	DMP	**0.78**	0.59	0.29
MC	11	**PK**	M	**2.38**	0.00	0.00
MC	13	**PK**	DMP	**1.02**	0.42	0.41
MC	14	**PK**	GA	**1.43**	0.60	0.47
MC	15	**PK**	PA	**0.71**	0.62	0.27
MC	16	**PK**	PA	**0.89**	0.48	0.29
MC	17	**PK**	NS	**1.25**	0.56	0.55
MC	18	**PK**	M	**0.92**	0.48	0.37
MC	19	**PK**	NS	**0.72**	0.57	0.59
MC	20	**PK**	NS	**0.69**	0.54	0.33
MC	26	**PK**	DMP	**1.44**	0.68	0.36
MC	3	**PS**	GSS	0.56	**0.58**	0.42
MC	4	**PS**	NS	0.35	**0.53**	0.23
MC	6	**PS**	DMP	0.67	**1.19**	0.34
MC	7	**PS**	PA	0.82	**1.19**	0.31
MC	8	**PS**	NS	0.84	**0.96**	0.29
MC	9	**PS**	GSS	0.47	**0.59**	0.32
MC	12	**PS**	DMP	0.64	**0.97**	0.25
MC	21	**PS**	M	0.52	**0.89**	0.25
MC	22	**PS**	M	0.24	**0.41**	0.22
MC	23	**PS**	GSS	**1.15**	**1.18**	0.43
MC	24	**PS**	GSS	**0.87**	**0.85**	0.28
MC	25	**PS**	NS	**1.75**	**1.14**	0.56
MC	27	**PS**	DMP	0.82	**1.46**	0.25
MC	28	**PS**	NS	0.83	**1.75**	0.00

*These content strand classification is based on the assessment framework. DMP—Data Management and Probability; GSS—Geometry and Spatial Sense; M—Measurement; NS—Number Sense; PA—Patterning and Algebra

4.2.5 From ability dimensions to reference composites as mathematical constructs

As explained in the Methods chapter, the direction of the reference composite (RC) is determined by an eigenvector that corresponds to the largest eigenvalues of each **a'a** matrix. From this eigenvector, angles from each ability coordinate axis were determined. These three angles indicate the direction of reference composite in the ability space. Each RSD has three eigenvectors corresponding to the three RCs. The angles of each RC for all RSDs—five 2006 RSDs and five 2005A RSDs—are presented in Tables 4.30 to 4.32. While there are some variations across RSD sets, three angles of each RC for each RSD are similar across RSD sets. The similarity of angles of each RC across RSDs indicates that each RC across the five RSD sets measures a similar combination of abilities. The angles of RCs for both tests will be very similar if the two tests measure the same construct and were built on the same test specification. As shown in Table 4.29, for example, there is some discrepancy in angles between the 2006 test and the 2005 test. This is an expected result because the two tests were developed according to different test specifications. However, the direction of each RC in the 2006 test is relatively close to that in the 2005 test, allowing comparison of abilities on three constructs. If two tests have the same constructs and the same test specification, the direction of RCs are expected to be similar. This suggests that test equating applying RC approach will work well.

Table 4.30
Angles for reference composite _PK

2006_RSD					2005A_RSD				
601	602	603	604	605	501	502	503	504	505
21.02	21.31	23.37	22.27	24.35	26.53	26.12	26.94	23.82	28.94
73.86	74.63	70.49	70.37	71.03	74.19	73.02	72.01	72.88	73.81
76.89	75.60	77.64	79.92	75.30	69.27	70.76	70.64	73.95	66.70

Table 4.31
Angles for reference composite 2_PS

2006_RSD					2005_RSD				
601	602	603	604	605	501	502	503	504	505
49.95	50.42	54.64	52.24	49.85	48.45	45.34	42.18	49.57	42.41
42.10	41.41	37.53	40.15	42.19	45.27	48.55	51.59	43.45	51.40
79.17	79.76	79.02	78.35	79.18	75.27	74.92	75.24	76.78	75.16

Table 4.32
Angles for reference composite 3_C&R

2006_RSD					2005_RSD				
601	602	603	604	605	501	502	503	504	505
59.90	60.97	61.24	54.94	61.26	59.57	60.40	59.88	60.82	64.20
65.12	66.08	62.77	66.67	65.91	67.96	65.75	70.87	69.88	70.80
40.90	39.22	41.60	44.24	39.11	39.08	39.97	36.82	36.64	33.06

From the eigenvectors of **a′a**, rotation matrixes for each RC were identified using the procedures explained in the Methods chapter. Each RSD has three reference composites. Therefore, for each RSD, there were three rotation matrixes to make ability coordinate axes aligned with three reference composite coordinates after being rotated. Rotation matrixes used for this study are presented in Tables 4.33 and 4.34.

Table 4.33
2005A rotation matrices

	RC1_PK			RC2_PS			RC3_C&R		
RSD501	0.895	-0.291	-0.339	0.663	-0.728	-0.174	0.507	-0.595	-0.624
	0.273	0.957	-0.103	0.704	0.686	-0.185	0.375	0.804	-0.462
	0.354	0	0.935	0.254	0	0.967	0.776	0	0.630
RSD502	0.898	-0.309	-0.313	0.703	-0.686	-0.189	0.494	-0.639	-0.589
	0.292	0.951	-0.102	0.662	0.728	-0.178	0.411	0.769	-0.49
	0.330	0	0.944	0.260	0	0.966	0.766	0	0.642
RSD503	0.891	-0.328	-0.313	0.741	-0.643	-0.195	0.502	-0.547	-0.670
	0.309	0.945	-0.109	0.621	0.767	-0.164	0.328	0.837	-0.438
	0.332	0	0.943	0.255	0	0.967	0.801	0	0.599
RSD504	0.915	-0.306	-0.263	0.649	-0.746	-0.152	0.488	-0.577	-0.656
	0.294	0.952	-0.085	0.726	0.666	-0.171	0.344	0.817	-0.463
	0.276	0	0.961	0.229	0	0.974	0.802	0	0.597
RSD505	0.875	-0.304	-0.377	0.738	-0.646	-0.196	0.435	-0.603	-0.669
	0.279	0.953	-0.120	0.624	0.764	-0.165	0.329	0.798	- 0.505
	0.396	0	0.919	0.256	0	0.967	0.838	0	0.546

Table 4.34
2006 Rotation Matrices

	RC1_PK			RC2_PS			RC3_C&R		
RSD601	0.933	-0.285	-0.218	0.644	-0.756	-0.123	0.502	-0.643	-0.579
	0.278	0.958	-0.065	0.742	0.655	-0.142	0.421	0.766	-0.486
	0.227	0	0.974	0.188	0	0.982	0.756	0	0.655
RSD602	0.932	-0.274	-0.239	0.637	-0.762	-0.115	0.485	-0.641	-0.595
	0.265	0.9618	-0.068	0.75	0.647	-0.136	0.405	0.768	-0.497
	0.249	0	0.969	0.178	0	0.984	0.775	0	0.632
RSD603	0.918	-0.342	-0.201	0.579	-0.808	-0.112	0.481	-0.689	-0.542
	0.334	0.940	-0.073	0.793	0.590	-0.154	0.458	0.725	-0.515
	0.214	0	0.977	0.191	0	0.982	0.748	0	0.664
RSD604	0.925	-0.341	-0.165	0.612	-0.780	-0.126	0.574	-0.568	-0.590
	0.336	0.94	-0.090	0.764	0.625	-0.158	0.396	0.823	-0.407
	0.175	0	0.9850	0.202	0	0.979	0.716	0	0.698
RSD605	0.911	-0.336	-0.239	0.645	-0.754	-0.123	0.481	-0.648	-0.592
	0.325	0.942	-0.085	0.741	0.656	-0.142	0.408	0.762	-0.502
	0.254	0	0.967	0.188	0	0.982	0.776	0	0.631

By post-multiplying ability matrixes by the rotation matrixes presented below, theta values on ability coordinate axes were translated into values on the rotated coordinate axes which are aligned with each reference composite, i.e., each construct. Using the same rotation matrixes, theta matrixes can be rotated so that the values are aligned with the three RC coordinate axes.

4.2.6 Abilities on constructs as reference composites

The descriptive statistics—mean, standards deviation, minimum, and maximum—for θ coordinates on the three-dimensional ability space before rotation onto RC coordinate axes are reported in Table 4.35.

Table 4.35
Mean, standard deviation, minimum and maximum of θ vectors before rotation

RSD		Mean	SD	Min	Max	RSD		Mean	SD	Min	Max
501A	θ_1	-0.043	0.818	-2.556	2.239	601	θ_1	0.003	0.834	-3.017	2.227
	θ_2	0.080	0.607	-2.022	1.698		θ_2	0.003	0.754	-2.86	1.941
	θ_3	-0.010	0.216	-0.709	0.792		θ_3	0.004	0.766	-3.302	2.064
502A	θ_1	-0.039	0.818	-2.59	2.244	602	θ_1	0.001	0.834	-2.768	2.209
	θ_2	0.082	0.632	-2.135	1.855		θ_2	0.002	0.763	-2.615	1.904
	θ_3	0.073	0.223	-0.663	0.882		θ_3	0.001	0.774	-3.284	2.16
503A	θ_1	-0.044	0.808	-2.499	2.213	603	θ_1	0.003	0.824	-3.021	2.175
	θ_2	0.115	0.597	-1.802	1.722		θ_2	0.004	0.776	-2.996	1.96
	θ_3	0.045	0.248	-0.795	0.881		θ_3	0.004	0.760	-3.3	2.084
504A	θ_1	-0.025	0.830	-2.56	2.293	604	θ_1	0.006	0.826	-2.936	2.241
	θ_2	0.093	0.602	-1.891	1.685		θ_2	0.007	0.765	-2.682	1.908
	θ_3	-0.068	0.203	-0.663	0.603		θ_3	0.006	0.751	-3.201	1.995
505A	θ_1	-0.029	0.865	-2.898	2.409	605	θ_1	0.000	0.829	-3.004	2.206
	θ_2	0.106	0.569	-1.893	1.618		θ_2	0.001	0.760	-2.859	1.858
	θ_3	-0.037	0.230	-0.828	0.782		θ_3	0.000	0.768	-3.394	2.083

There are two important observations to be addressed from Table 4.35. First, the means of $\theta_{1,}$ $\theta_{2,}$ and θ_3 across all 2006 RSDs are very close to zero because each RSD set was calibrated separately. When the 2005 test was linked to the 2006 test by the FCIP linking method, the mean of θ_1 of 2005 students was a little lower than that of 2006 students in all five RSD sets. In the case of θ_2, mean thetas of 2005 students were a little higher than those of 2006 students in all five RSD sets. The difference in θ_3 was not consistent across the five 2005 RSD sets. While the mean of θ_3 of RSD502A and RSD503A was higher than in their paired 2006 RSD sets—RSD602 and RSD603 each—those of RSD501A, RSD 504A and RSD505A were a little lower than in 2006.

To check if these differences are statistically significant, t-tests were conducted. All differences were statistically significant. However, the statistical significance might be due to the large sample size—10,000. So, Cohen's *d* was computed for each θ to check effect size for the difference between each paired θs, using the pooled variance because variances were not equal. The results are reported in 4.36. As shown in Table 4.36, all effect sizes were negligible, suggesting little difference in abilities between the 2005 and the 2006 students.

Another important point is that standard deviations of θ in the 2005 test data are much smaller than those in the 2006 test data, especially in Dimension 3. This means that the standard deviations shrank when the 2005 test data were linked to the 2006 test data using FCIP, which was also observed in the UIRT FCIP linking.

Table 4.36
Effect size of the difference in θs: Cohen's *d*

RSD Pair by θ	Effect size
θ1_ 501a1 - 601a1	-0.025
θ2_ 501a2 - 601a2	0.02
θ3_ 501a3 - 601a3	-0.005
θ1_ 502a1 - 602a1	0.011
θ2_ 502a2 - 602a2	0.042
θ3_ 502a3 - 602a3	0.071
θ1_ 503a1 - 603a1	0.007
θ2_ 503a2 - 603a2	0.043
θ3_ 503a3 - 603a3	0.048
θ1_ 504a1 - 604a1	-0.024
θ2_ 504a2 - 604a2	0.025
θ3_ 504a3 - 604a3	-0.047
θ1_ 505a1 - 605a1	-0.011
θ2_ 505a2 - 605a2	0.033
θ3_ 505a3 - 605a3	-0.01

To produce a full θ matrix aligned with three reference composite coordinates, each RSD set needs to be rotated three times using three rotation matrixes for three reference composites. Applying the rotation matrixes reported above, three θ coordinates specifying the location of each student in the θ space were translated into the values on reference composite coordinate axes. When θ coordinates were translated into the θ values on construct coordinate axes (Table 4.37 and table 4.38), the pattern of difference changed reflecting the fact that locations in the rotated ability space do not correspond to locations identified in the ability space before rotation. Overall, θ values on constructs are a little higher in 2005 than in 2006, suggesting different results in proficiency classification change than in the UIRT linking case. Shrinkage of standard deviations in the 2005 test data also occurred in construct abilities, but the degree of shrinkage is smaller than in ability dimensions.

Table 4.37 Mean, standard deviation, minimum and maximum of θ vectors after rotation: 2005

		Mean	SD	Min	Max
RSD501	RC_θ_1	-0.020	0.8767	-2.811	2.462
	RC_θ_2	0.026	0.9178	-3.099	2.663
	RC_θ_3	0.001	0.6308	-1.995	1.769
RSD502	RC_θ_1	0.013	0.8843	-2.854	2.524
	RC_θ_2	0.046	0.9312	-3.159	2.727
	RC_θ_3	0.070	0.6241	-1.935	1.834
RSD503	RC_θ_1	0.012	0.8596	-2.633	2.441
	RC_θ_2	0.050	0.9121	-2.838	2.661
	RC_θ_3	0.052	0.5414	-1.496	1.528
RSD504	RC_θ_1	-0.014	0.9230	-2.947	2.620
	RC_θ_2	0.036	0.9284	-3.073	2.713
	RC_θ_3	-0.035	0.6437	-2.039	1.795
RSD505	RC_θ_1	-0.011	0.8924	-3.019	2.524
	RC_θ_2	0.035	0.9426	-3.292	2.747
	RC_θ_3	-0.009	0.5551	-1.789	1.516

Table 4.38 Mean, standard deviation, minimum and maximum of θ vectors after rotation: 2006

		Mean	SD	Min	Max
RSD601	RC_θ_1	0.004	1.0841	-4.360	3.048
	RC_θ_2	0.005	1.1453	-4.684	3.175
	RC_θ_3	0.006	1.1922	-5.213	3.436
RSD602	RC_θ_1	0.002	1.0852	-4.065	3.020
	RC_θ_2	0.003	1.1383	-4.274	3.127
	RC_θ_3	0.002	1.1829	-4.888	3.378
RSD603	RC_θ_1	0.005	1.0943	-4.480	3.048
	RC_θ_2	0.006	1.1396	-4.753	3.126
	RC_θ_3	0.007	1.1947	-5.292	3.385
RSD604	RC_θ_1	0.009	1.0813	-4.134	3.038
	RC_θ_2	0.010	1.1491	-4.439	3.192
	RC_θ_3	0.010	1.1971	-4.958	3.429
RSD605	RC_θ_1	0.001	1.1107	-4.516	3.127
	RC_θ_2	0.001	1.1462	-4.666	3.172
	RC_θ_3	0.001	1.1825	-5.231	3.420

4.3 MIRT Approach to Proficiency Classification – Compensatory vs. Conjunctive

In the compensatory approach, a composite score as the equally weighted average of three construct ability scores expressed as θ was computed for proficiency classification. The cut scores for the 20th percentile in the 2005 RSD sets were -.698, -.668, -.635, -.729, and -.879 respectively (Table 4.39). When these cut scores were applied to the 2006 RSD sets matched with each 2005 RSD set, the proficiency rates in 2006 were 73.33%, 72.69%, 71.94%, 74.61%, and 78.18% each (Table 4.39). With 74.15% in average across the five paired RSD sets, proficiency rate decreased by about 6% in 2006 than in 2005. MIRT FCIP linking suggested that 2005 students performed better than 2006 students, which is different from the UIRT FCIP linking results. χ^2 test results showed that the difference in the proficiency rate is statistically significant for all five RSD sets (Table 4.40).

Table 4. 39 Cut-point at the 20th percentile in 2005 and proficiency rate: 2006

	Cut-point	2006 Proficiency rate
501a	***-0.698***	73.33%
502a	***-0.668***	72.69%
503a	***-0.635***	71.94%
504a	***-0.729***	74.62%
505a	***-0.879***	78.18%
Average		74.15%

Table 4.40 χ^2 results_MIRT_compensatory_20th percentile

	χ^2
501-601	124.34
502-602	147.95
503-603	177.93
504-604	82.50
505-605	10.01

When the same procedure was applied to the 50th percentile score in 2005 as a cut-point, the change of the proficiency rates between 2005 and 2006 decreased dramatically, showing almost similar performance between 2005 students and 2006 students (Table 4.41). Therefore, the discrepancy in proficiency rate difference between 2005 and 2006 decreased when the 50th percentile cut-point was applied. The results of χ^2 tests confirmed that there is no statistically significant difference in the proficiency rate between 2005 and 2006 (Table 4.42).

Table 4. 41 Cut-point at the 50th percentile in 2005 and proficiency rate: 2006

	Cut-point	2006 Proficiency rate
501a	***0.018***	50.79%
502a	***0.067***	49.00%
503a	***0.062***	49.43%
504a	***0.016***	51.15%
505a	***0.017***	50.76%
Average		50.23%

Table 4.42 χ^2 results_MIRT_compensatory_50th percentile

	χ^2
501-601	1.25
502-602	2.00
503-603	0.65
504-604	2.65
505-605	1.16

In the conjunctive approach to proficiency classification, the basic idea is that proficiency level is set for each construct separately and students need to be proficient at all of the three constructs to be classified to be proficient. The same two cut points, the 20th percentile and the 50th percentile, were used as a cut-point for each construct.

Therefore, overall proficiency rates in 2005 are lower than 50% when the 50th percentile cut-point for each construct is applied and lower than 80% when the 20th percentile cut-point for each construct is applied. To determine the proficiency rate in 2005 when cut-points at the 20th percentile and at the 50th percentile for each construct are applied, first cut scores for each construct at the 20th percentile and at the 50th percentile for the five 2005 RSD sets were identified. The cut-points for each construct at the 20th percentile and the 50th percentile are presented in Table 4.43. As shown in the table, the cut-points for the 20th percentile are somewhat similar, but those for the 50th percentile across five RSD sets vary a little.

Table 4.43 Cut-scores for each construct at the 20th and the 50th percentile in the conjunctive approach: 2005

	The 20th percentile			The 50th percentile		
	RC1	RC2	RC3	RC1	RC2	RC3
RSD501	-0.7864	-0.7688	-0.5579	-0.0046	0.0367	0.0156
RSD502	-0.7612	-0.7604	-0.4828	0.0392	0.0609	0.0874
RSD503	-0.7478	-0.7388	-0.4432	0.0318	0.0721	0.0666
RSD504	-0.8257	-0.7763	-0.6059	0.0075	0.0467	-0.0207
RSD505	-0.8028	-0.7897	-0.519	0.0107	0.0516	0.0064

For comparison, the cut-points for each construct at the 20th percentile and the 50th percentile for the five 2006 RSD sets are presented in Table 4.44. As shown in the table, the two cut-points, the 20th percentile and the 50th percentile, across five RSD sets are very similar. While the means of construct ability were similar between 2005 and 2006 (Tables 4.37 and 4.38), the scores for each cut-point for each construct across the five RSD sets are somewhat different between 2005 and 2006.

Table 4.44 Cut-scores for each construct at the 20th and the 50th percentile in the conjunctive approach: 2006

	The 20th percentile			The 50th percentile		
	RC1	RC2	RC3	RC1	RC2	RC3
RSD601	-0.9402	-0.9828	-1.0082	0.0108	0.053	0.06
RSD602	-0.9221	-0.9603	-0.9847	0.0141	0.0481	0.0579
RSD603	-0.925	0.948	-1.0009	0.0179	0.0554	0.0564
RSD604	-0.9207	-0.9587	-1.0065	0.024	0.0662	0.059
RSD605	-0.9375	-0.9518	-0.9966	0.0201	0.0494	0.0518

As already explained, the same cut scores for the 2005 RSD sets were applied to the matched 2006 RSD sets to obtain proficiency rates in 2006. The proficiency rates at the 20th percentile and the 50th percentile when the 2005 cut-points were applied to the 2006 test results are presented in Table 4.45 and Table 4.46. When the 20th percentile cut-point for each construct was applied, the proficiency rates in 2005 across the five RSD sets were about 77% in 2005 and 65% to 69% in 2006, resulting in an 8% to 12% decrease in proficiency rate from 2005 to 2006.

Table 4.45 Proficiency rates at the 20th percentile: conjunctive classification

2005 and 2005 RSD pair	501-601	502-602	503-603	504-604	505-605
2005	77.04%	77.24%	76.84%	76.91%	76.58%
2006	67.16%	65.61%	64.77%	69.06%	66.71%
Change (2006-2005)	**-9.88%**	**-11.63%**	**-12.07%**	**-7.85%**	**-9.87%**

When the 50th percentile cut-point for each construct was applied in the conjunctive approach, the proficiency rates in 2005 across the five RSD sets were about 44% to 45% in 2005 and 43% to 46% in 2006 (Table 4.46). The proficiency rate change from 2005 to 2006 was dramatically smaller compared to the 20th percentile. Except the RSD502-RSD602 pair, 2006 showed little difference in proficiency rate.

Table 4.46 Proficiency rates at the 50th percentile: conjunctive classification

2005 and 2005 RSD pair	501-601	502-602	503-603	504-604	505-605
2005	44.94%	45.49%	44.05%	45.00%	44.44%
2006	45.21%	42.97%	44.09%	46.23%	45.13%
Change (2006-2005)	0.27%	-2.52%	0.04%	1.23%	0.69%

To check if the percentage change for each pair is statistically significant, χ^2 tests were conducted. The result is reported in Table 4.47. The result shows that increase in proficiency rate is statistically significant for only one RSD set.

Table 4.47 χ^2 results_MIRT_conjunctive_50th percentile

	χ^2
501-601	0.15
502-602	**13.87**
503-603	0.00
504-604	3.05
505-605	0.96

In the conjunctive approach to proficiency classification, students need to be proficient at each construct being measured in order to be classified as proficient. Because failing to achieve proficiency in any one of constructs results in "non-proficiency" classification, the percentage of students who failed to achieve overall proficiency when they are proficient in one or two constructs were also examined. The relative percentages of students who passed cut-points in zero, one, two, and all of three constructs are presented in Table 4.48 (2005) and Table 4.49 (2006).

Table 4.48 Relative percentage by the number of proficient constructs: 2005

		RSD501	RSD502	RSD503	RSD504	RSD505
20th percentile	pass 0	17.07	17.12	16.90	16.91	16.64
	pass 1	2.94	3 .03	3.04	3.09	3.30
	pass 2	2.92	2.61	3.22	3.09	3.48
	pass all	77.07	77.24	76.84	76.91	76.58
50th percentile	pass 0	45.05	45.57	43.99	45.02	44.32
	pass 1	4.85	4.37	6.07	4.96	5.80
	pass 2	5.16	4.57	5.89	5.02	5.44
	pass all	44.94	45.49	44.05	45.00	44.44

Table 4.49 Relative percentage by the number of proficient constructs: 2006

		RSD601	RSD602	RSD603	RSD604	RSD605
50th percentile	pass 0	21.62	21.34	22.02	20.84	21.33
	pass 1	4.09	4.94	4.48	3.58	3.86
	pass 2	7.13	8.11	8.73	6.52	8.10
	pass all	67.16	65.61	64.77	69.06	66.71
20th percentile	pass 0	43.67	44.51	45.09	43.39	43.97
	pass 1	6.00	6.27	5.44	5.49	5.60
	pass 2	5.12	6.25	5.38	4.89	5.30
	pass all	45.21	42.97	44.09	46.23	45.13

4.4 Comparison of Proficiency Rates Change between UIRT Approach, MIRT Compensatory Approach, and MIRT Conjunctive Approach

As already discussed, to link the two years' test data—2005 and 2006—the FCIP linking method was employed. When the 20th percentile score was used as a cut-point for proficiency, the three approaches to proficiency classification produced different results (Table 4.48). In the UIRT linking approach, there was little change in proficiency rate between 2005 and 2006, with a 0.1% decrease in proficiency rate in 2006 as compared to 2005. In the MIRT compensatory approach, there was 5.85% decrease in proficiency rate in 2006 as compared to 2005. In the MIRT conjunctive approach, there was an even larger decrease of 10.26% in proficiency rate in 2006 as compared to 2005. When the conjunctive approach is applied, it is harder to achieve proficiency because students must be proficient on all of the three constructs.

Table 4.50 Comparison of proficiency rate change (%) by proficiency classification approach at the 20th percentile

Proficiency Classification approach		*501 -601*	*502 -602*	*503 -603*	*504-604*	*505-605*	*mean*
UIRT	2005 Proficiency rate	80.00	80.00	80.00	80.00	80.00	80.00
	2006 Proficiency rate	79.93	79.82	79.41	80.15	80.17	79.90
	Change (2006-2005)	***-0.07***	***-0.18***	***-0.59***	***0.15***	***0.17***	***-0.10***
MIRT compensatory	2005 Proficiency rate	80.00	80.00	80.00	80.00	80.00	80.00
	2006 Proficiency rate	73.33	72.69	71.94	74.62	78.18	74.15
	Change (2006-2005)	***-6.67***	***-7.31***	***-8.06***	***-5.38***	***-1.82***	***-5.85***
MIRT conjunctive	2005 Proficiency rate	77.04	77.24	76.84	76.91	76.58	76.92
	2006 Proficiency rate	67.16	65.61	64.77	69.06	66.71	66.66
	Change (2006-2005)	**-9.88**	**-11.63**	**-12.07**	**-7.85**	**-9.87**	***-10.26***

Table 4.51 Comparison of proficiency rate change (%) by proficiency classification approach at the 50th percentile

Proficiency Classification approach		*501 -601*	*502 -602*	*503 -603*	*504-604*	*505-605*	*mean*
UIRT	2005 Proficiency rate	50.00	50.00	50.00	50.00	50.00	50.00
	2006 Proficiency rate	*51.53*	*51.13*	*50.70*	*51.22*	*50.98*	51.11
	Change (2006-2005)	***1.53***	***1.13***	***0.70***	***1.22***	***0.98***	***1.11***
MIRT compensatory	2005 Proficiency rate	50.00	50.00	50.00	50.00	50.00	50.00
	2006 Proficiency rate	*50.79*	*9.00*	*49.43*	*51.15*	*50.76*	50.23
	Change (2006-2005)	***0.79***	***-1.00***	***-0.57***	***1.15***	***0.76***	***0.23***
MIRT conjunctive	2005 Proficiency rate	44.94	45.49	44.05	45.00	44.44	44.78
	2006 Proficiency rate	45.21	42.97	44.09	46.23	45.13	44.73
	Change (2006-2005)	***0.27***	***-2.52***	***0.04***	***1.23***	***0.69***	***-0.06***

When the 50th percentile score was set as the cut-score for proficiency, the three different approaches to proficiency classification produced little the difference (Table 4.49). For the UIRT approach, there was a 1.11% increase in proficiency rate in 2006 as compared to 2005, 0.23% of an increase for the MIRT compensatory approach, and little change for the MIRT conjunctive approach. The difference in proficiency rate change between the UIRT approach and the MIRT compensatory approaches can be explained as follows.

In the UIRT approach, it is assumed that both tests are measuring a unidimensional construct or a composite of the same constructs. As Wang (1985) showed, when there are multiple ability dimensions, i.e., when a test is measuring multiple constructs, the composite of abilities approximated by the UIRT approach is determined by the statistical estimation algorithm. A single measure as the best combination of multiple construct abilities is determined by a given statistical criterion such as least squares in regression. In other words, the single number reported for ability is a composite of multiple abilities in the UIRT approach. It is a statistical composite that does not support substantive interpretation of the reported score.

In contrast, for the MIRT approach, constructs are identified by considering substantive meanings being measured by each item. In the MIRT approach, statistical analysis of dimensionality—MIRT calibration—the examination of dimensional structure through hierarchical cluster analyses, the identification of substantial meaning of ability dimensions statistically identified, and the identification of meaningful constructs in the context of subject area being assessed make possible measuring construct abilities with substantial meanings.

The MIRT approach to proficiency classification requires the generation of individual ability scores for each construct before calculating a single compensatory composite score. Because there are separate scores for each construct, it is possible to get the norm-referenced interpretation of performance on each construct measured. In the compensatory approach, it is not possible to know what the score at each construct means in the terms of criterion-referenced interpretation. When the conjunctive approach is applied, however, it allows criterion-referenced interpretation for each construct.

CHAPTER 5

SUMMARY, DISCUSSION, AND CONCLUSION

5.1 Summary of the Research

The major purpose of this study was to show educational policy implications of psychometric decisions on educational measurement by exploring the effects of selecting different linking approaches – UIRT linking vs. MIRT linking – on proficiency rate changes across years. The result shows that different linking approaches and different choices of proficiency classification models as a result of selecting different linking approaches produce different conclusions on the educational progress inferred from increased or decreased proficiency rates across years.

This study had two additional purposes. First, this study intended to explore the feasibility of alternative approach to subscore reporting practices by identifying what typical large-scale mathematics assessment tests are measuring. By conducting MIRT dimensionality analyses of real data from a large-scale mathematics achievement assessment program, this study identified three meaningfully interpretable constructs – procedural knowledge, problem solving, and communication and representation. By applying psychometric properties of the reference composite concepts to measuring abilities on constructs identified through MIRT cluster analysis rather than on statistically determined ability dimensions, this study documented the feasibility of measuring abilities on educationally meaningful constructs. Measuring construct level abilities provides an alternative approach to subscore reporting and to tracing growth of mathematical abilities measured on construct level.

Second, this study was designed to show the feasibility of conducting MIRT linking for mixed format test forms through the analyses of real data. Some researchers such as Traub (1993) have argued that different test items measure different constructs. This study confirmed the argument through MIRT cluster analysis. By identifying which item measures which construct, this study showed that constructed-response items measure primarily the communication and representation construct and multiple-choice items measure mainly procedural knowledge and problem solving. However, the distinction is not absolute because there were several multiple-choice items identified as measuring the communication and representation construct. This suggests that test format constrains constructs intended to be measured at a certain level. When constructed-response items measure different constructs which are difficult to measure through multiple-choice items, mixed format tests are desirable to measure a full range of important constructs. However, MIRT linking has been typically conducted using only multiple choice items because of difficulty with conducting MIRT linking of mixed format test forms. This study showed the feasibility and thus practical applicability of MIRT linking of mixed format tests.

In this study, a fixed common item linking (FCIP) method was selected to link two years' test data in both the UIRT and MIRT linking approaches. Five random sample data (RSD) sets of 10,000 samples were selected for the years 2005 and 2006 and each RSD set of one year (2005) was linked to its matched RSD set of 2006. For an UIRT approach, the PARSCALE program was employed. The two years' test data were calibrated separately by running PARSCALE and 2005 results were recalibrated by fixing the common items with the item parameters calibrated in 2006 data. For MIRT linking, each RSD set was calibrated by running the BMIRT Program. To estimate abilities for 2005 by fixing the common items as the item parameters calibrated in 2006 test data, the BMIRTanchor program was employed.

The specific procedures of conducting MIRT linking mixed format tests selected in this study are as follows. First, five RSD sets from each year's test data were selected using the MATLAB program. Second, each RSD was calibrated separately by running BMIRT. Each 2005 RSD set was recalibrated using BMIRTanchor by fixing the common items with item parameters calibrated from the matched 2006 RSD. Third, MIRT cluster analysis was conducted on each RSD set using MATLAB. Hierarchical cluster analysis (HCA) using the Ward method was employed. Based on the results of HCA and the loading structure of discrimination parameters, each item was assigned to one of three clusters identified through HCA. Then, item review was conducted through an item review committee activity. In the review committee, each item was reviewed by the committee and the construct measured in common by each cluster was identified. Fifth, reference composite (RC) for each construct was identified using the MATLAB program. Each reference composite identified indicated the direction of the best measurement in common of the items in each construct. From the direction of each reference composite, angles of each RC from each dimension were computed and the rotation matrix to align the coordinate axis of the dimensions with each construct, i.e., the reference composite, was obtained using MATLAB. By post-multiplying ability estimated on BMIRT (2006) and on BMIRTanchor (2005) by the rotation matrix, construct abilities were computed.

After conducting UIRT linking and MIRT linking, proficiency rate change was explored. Two cut-points for proficiency decision were selected – the 20th percentile and the 50th percentile, assuming that 80% of the students and 50% of the students in 2005 achieved proficiency in each case. In the UIRT approach, a single cut-score selected from the 2005 results at each cut-point was applied to identify the rate of proficiency in each matched 2006 RSD for each cut-point. In the MIRT approach, two classification approaches were employed – compensatory and conjunctive. In the compensatory approach, a single composite ability score for each student was calculated by applying the pre-determined weight of each construct. The cut-point scores at the 20th percentile and the 50th percentile were identified from the distribution of the composite ability scores. In the conjunctive approach, two cut-point scores at the 20th percentile and the 50th percentile for each construct were identified from the separate ability distribution by construct – resulting in three cut-point scores for each classification criterion. Three separate cut-point scores were applied to obtain the proficiency rate for both years' students. After conducting proficiency rate decision for each RSD per year for all three classification approaches, the proficiency rate changes were compared.

5.2. Findings and Discussions

The major findings from this study and discussions are provided below.

First, three meaningfully interpretable constructs were identified. The three constructs identified can be interpreted as mathematical constructs at the process level. These constructs do not correspond to content area or content strand. While instruction plan/schedules are made according to the content area to be covered, there are common mathematical abilities to be expected for students to learn at the process levels across content areas, as specified in the EQAO assessment framework or in NCTM Standards. It is possible to identify more minute constructs by dividing the test items into more clusters. This means that the direction of the best measurement of each item in each cluster is more similar to each other. The direction of the reference composite calculated from these items is closer to each item, which means less measurement error in projecting the best measurement of each item into the direction of the best measurement of a reference composite. This in turn suggests that there would be more measurement error involved in UIRT approach to calibration and linking.

Second, this study confirmed the findings from previous research that constructs measured by constructed-response items are different from those by multiple-choice items. In this study, it was found that constructed-response items measure mainly the communication and representation construct and that multiple-choice items measure mainly the procedural knowledge and problem solving constructs. However, this does not suggest that it is not possible to measure a certain construct by a certain item type. This study identified multiple-choice items which measured the communication and representation construct. However, it seems that item type constrains the scope and range of measurement. In terms of the amount of information by item type, it was found that constructed-response items provide more information on average than multiple-choice items. However, there were a couple of multiple-choice items which were comparable to the best constructed-response items in terms of providing information. Constructed-response items on the test studied here have more information for low ability students in general. The lack of information about low ability students available from multiple-choice items is partially due to the guessing parameter. This problem might be reduced when the guessing parameter can be modeled as a function of ability[29], which is now being tried (Martin, del Pino, & De Boeck, 2006).

Third, this study found that different linking approaches and different decision approaches to proficiency classification produced different results. Overall, the UIRT approach was favorable to the 2006 students. While there was little change in proficiency rate between 2005 and 2006 using the UIRT approach, both the MIRT compensatory and conjunctive approaches resulted in a decreased proficiency rate in 2006 compared to 2005 when the 20th percentile classification criterion was applied. When the 50th percentile classification criterion was applied the change in proficiency rate between two years was not statistically significant. This result strongly suggests the importance of selecting a linking method and a proficiency classification approach when evaluating educational progress by the change of proficiency rate.

5.3. Educational Policy Implications and Conclusion

This study explored the effects of linking method and proficiency classification criterion on the proficiency rate and thus on the evaluation of educational progress. The results of this study suggest several policy implications in relation to the current assessment policy, mathematics education, and educational measurement.

In relation to current assessment policy, three points can be made. First, this study raises a question of the current practice of reporting sub-scores. Currently most of the states reporting sub-score results report them by percent correct scores per content area, for example, per content strand such as algebra in the case of mathematics. This study shows that abilities required to solve the questions within a common content areas are not

[29] One example of those efforts is shown in Martin et al. (2006).

the same. For example, out of eight items in Data Management and Probability (DMP) content strand, three items measure the PS construct (problem solving), two measure the PK construct (procedural knowledge), and two measure the C&R construct (communication and representation). This means that abilities which students employ when they solve the problems in the DMP content strand are multiple, raising a question of the meaning of the percent correct score in DMP area.

The MIRT approach selected in this study presents a new possible approach to sub-score reporting. The previous MIRT approach to sub-score reporting (for example, Yao & Schwartz, 2006) focused on calculating sub-scores by content area by employing a MIRT confirmatory analysis. This approach can be interpreted basically as the content-perspective approach because items are assigned to dimensions as classified in test specifications assuming each content area is measuring a separate ability construct. Unlike the previous MIRT approach, the approach employed in this study can be interpreted as measurement-perspective approach. The focus of the approach is given to identifying constructs being measured in a test through both item content analysis and statistical analysis. This approach is more aligned with the basic idea of the multidimensional item response theory – each item needs more than one ability/skill. The MIRT confirmatory approach enforces each ability dimension represented by a content area on each item, which means each item belongs to one content area even when solving the item requires knowledge and skills across content areas. If this approach is applied to sub-score reporting, growth can be traced at each construct level and diagnostic information for instruction can be effectively obtained, especially when content area level achievement is available.

Second, the results of this study raise a question of the validity of the current UIRT linking approach to educational measurement by demonstrating the effects of linking method selected on measuring educational achievement. Depending on selected linking approach and proficiency classification approaches, inference on educational progress based on the test results change. When UIRT linking suggests improvement of educational achievement while MIRT linking results in little change in performance across years, which results should be considered to portray learning growth more accurately? Should we choose an approach which gives more "pleasant" results? Which linking approach is better from the statistical point of view, i.e., which approach is more accurate, can be identified by conducting a simulation study of linking and checking measurement error involved with linking, which was not tried in this study. However, which linking approach is better in terms of educational progress cannot be determined solely on the statistical basis Especially when the difference in the measurement error between the UIRT linking approach and MIRT linking is small, selection of a linking approach and proficiency classification approach should be made based on educational goals, which needs educational policy discourse.

Third, this study also documented the possible effects of standard setting on proficiency classification. While no standard setting procedure was involved in this study, it shows that different cut-points produced different results. Because the selection of a given cut-point can make a difference in the proficiency rate, great effort is required to ensure that a cut-point is psychometrically sound and substantially meaningful in terms of criterion-referenced evaluation.

In relation to mathematics assessment and education, this study showed that it is possible to identify educationally meaningful mathematical constructs to be measured. By providing a feasible method to measure ability on mathematical constructs, this study presents a way to measure learning growth in mathematical constructs. If subscores are reported by construct rather than content strand, it is possible to trace growth in mathematical learning at mathematical construct level. By allowing measurement and thus report of students' learning at construct level, this study presents a way to provide diagnostic information for mathematics teaching and learning. When mathematics learning is measured in terms of mathematical constructs, it is possible to identify that different assessments measure different mathematical constructs. In this case, mathematical achievement measured from a UIRT approach can be redefined. This study also suggests the limitation of typical standardized assessment consisting mainly multiple-choice items by showing that test format constrains mathematical constructs to be measured. This in turn raises a need to reconsider test development process in mathematics assessment. Mathematics assessment should be developed considering broad educational goals of mathematics education established through discourses and consensus among mathematics educators, policy makers, and other interested parts in society in general.

In relation to educational measurement and psychometrics, this study has several implications. First, this study provides a sound psychometric method to check if two test forms are measuring the same constructs. After linking item parameters from two (or more) separate calibrations though scale transformation or after conducting the FCIP linking or concurrent calibration, the direction of reference composite can be compared to determine how similar the direction of the best measurements of each corresponding reference composite is. This approach allows linking of multiple forms at the same time.

Second, the MIRT linking approach using reference composite also provides an effective and useful procedure to check construct shift and still measure growth on the same constructs. This is especially useful when tests are constructed based on the similar framework, but with different test specifications. If the directions of corresponding reference composites identified are similar, the test scores from different tests can be considered as comparable even though they are not constructed based on the same assessment framework. If tests are measuring different constructs, it is not possible to compare the results directly.

There are several areas which were not covered in this study, but deserve further research. First, this study employed the FCIP MIRT linking, which has not been tried often. Most MIRT linking has been conducted and studied using the oblique Procrustes method (Gower & Dijksterhuis, 2004). Comparing two methods through a simulation study will provide valuable information in evaluating psychometric qualities of two approaches to MIRT linking.

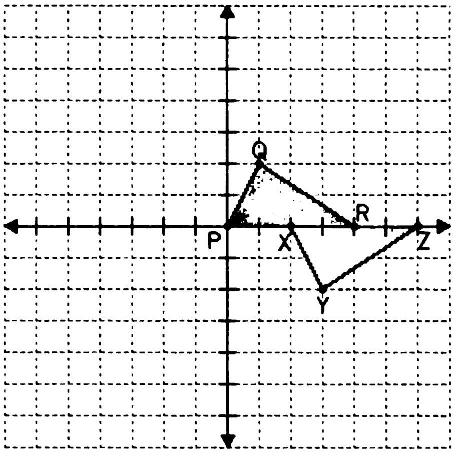
Second, this study was applied to the "calibration" level of linking. By applying the MIRT linking approach using reference composites, as in this study, to different levels of linking, it will be possible to present richer information on what tests are measuring and whether two scores from different tests are comparable.

Third, linking for this study was conducted using the 2006 test calibration as reference data. If linking was done from the opposite direction, the result might have been different. For the UIRT linking approach, direction might not be important. The comparison of proficiency classification using the entire data set shows a similar mismatch from both approaches. For MIRT linking, however, the choice of the reference year would have more effects especially when there was a substantial change in test specification such as test format. This was not studied in this research, but it deserves further research.

Psychometrics is not a perfect science. In applying psychometric theoretical procedures to practical testing situations, human judgments are always involved. Human judgment, however, should be based on sound scientific grounds, not political considerations or practical convenience. Because data structure constrains the possible approaches to linking or other psychometric applications, data collection design should be carefully developed considering the effects of the design on the test results and their implication to inferences made from the test results and policy decisions made based on the inferences.

Appendix A
EQAO Grade 6 Mathematics Test 2005

1. The triangle PQR has been transformed to the position XYZ.



What is a correct description of the transformation?

a translation to the right by 2 units and down by 2 units

b translation to the right by 4 units and reflection about the horizontal axis

c reflection about the horizontal axis followed by reflection about the vertical axis

d translation to the right by 2 units and reflection about the horizontal axis

2. Lan Ying and Sean love bike riding. They take off from the same place, at the same time, going in the same direction.

   Lan Ying rides at a steady speed of 6 km/h, and Sean rides at a steady speed of 4 km/h.

   How far apart will they be in 3 hours?

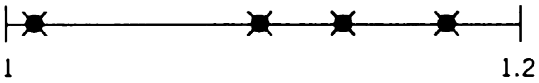
   a 6 km

   b 12 km

   c 18 km

   d 30 km

3. Which set of values is represented by the 4 points on the number line?



   a 1.05, 1.1, 1.5, 1.9

   b 1.01, 1.10, 1.13, 1.17

   c 1.1, 1.11, 1.15, 1.19

   d 1.01, 1.10, 1.12, 1.5,

4. Which pattern has this rule: decrease by subtracting the same amount from each term?

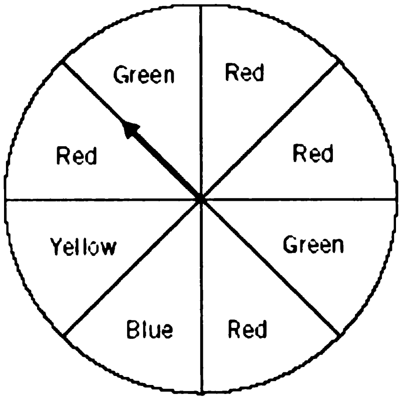
a 20, 10, 5, 2.5, 1.25

b 20, 18, 16, 14, 12

c 20, 25, 30, 40, 45

d 20, 40, 80, 160, 320

5. A circle is divided into sections of equal size.



What is the probability that the spinner will stop on red?

a 1 out of 4

b 3 out of 8

c 3 out of 4

d 1 out of 2

6. Write 207.083 in expanded form.

a 200 + 7 + 0.08 + 0.003

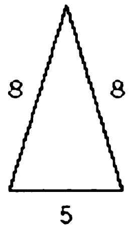
b 20 + 7 + 0.8 + 0.3

c 200 + 7 + 0.08 + 0.03

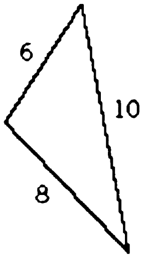
d 200 + 7 + 0.8 + 0.003

7. Which triangle must have at least one 60° angle?

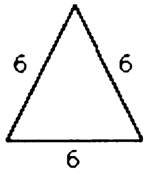
a



b

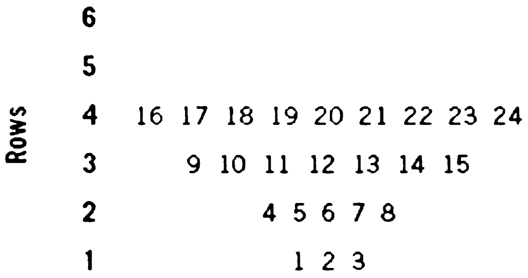


c



d none of the above

8. The seating plan diagram shows how the chairs need to be arranged for the school concert. The pattern continues to Row 15.



Which rule below can be used to find the number of seats in each row?

a There is an odd number of seats in each row.

b Each row has more seats.

c Each row has 2 more seats than the row before it.

d There are 24 seats.

9. Sarah types 115 words in 5 minutes. Mary types 174 words in 8 minutes.

- Who types faster?

Show your work.

__________ types faster.

10. Jessica wants to pour 5 kg of sugar into smaller bags.

- If each bag holds 250 grams, how many bags does she need?

Show your work.

Jessica needs _____ bags.

11. Jaspreet is wondering why she made the honour roll but her friend Cynthia did not. She knows that an overall percentage of 80 or higher makes the honour roll. She is not sure if the school uses the mean, median or mode to calculate the percentage.

Jaspreet's Marks

70 73 73 83 78 93 87 85 80

Cynthia's Marks

87 75 76 84 78 94 70 84 79

- Determine whether the school uses the mean, median or mode to calculate the overall percentage.

Show your work.

The school uses ______________ to calculate the overall percentage.

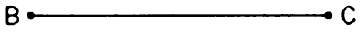
12. Follow the instructions below to create a polygon. You will need a protractor and a ruler. Start with the line BC below.

1) At Point B, use a protractor to create an angle of 30° with sides measuring 5 cm each.

2) Label it ∠ABC.

3) At Point C, create an angle of 150° and label it ∠BCD.

4) Connect Point D to Point A with a 5 cm line to complete the polygon.



What is the name of this type of polygon? ______________

13. Sharon works at the local gym. She must buy number stickers to label the lockers in the change rooms. Stickers are sold as individual digits. There are 79 lockers, which will be labelled 1 to 79.

- How many stickers of each digit should she buy?

To label locker 19, Sharon buys two stickers: 1 and 9.

Show your work.

14. Look at the figures and table of values below.

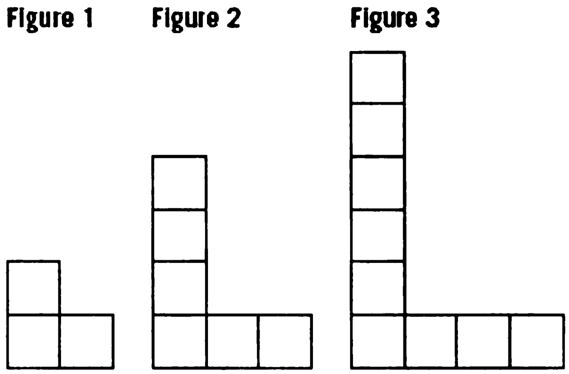


Figure	1	2	3	4	...	8	9
Squares	3	6	9	12	...	24	27
Perimeter	8	14	20	26	...	?	?

Which set of numbers represent the perimeter of the 8th and 9th figures?

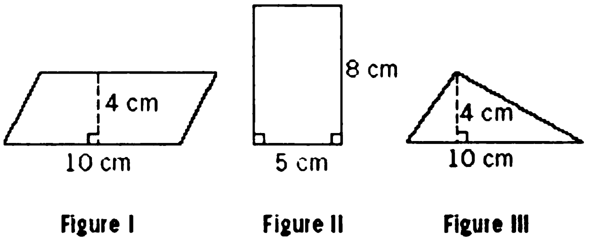
a 44, 50

b 48, 54

c 50, 56

d 68, 76

15. Look at the figures.



Which 2 figures have the same area?

a Figure I and Figure II

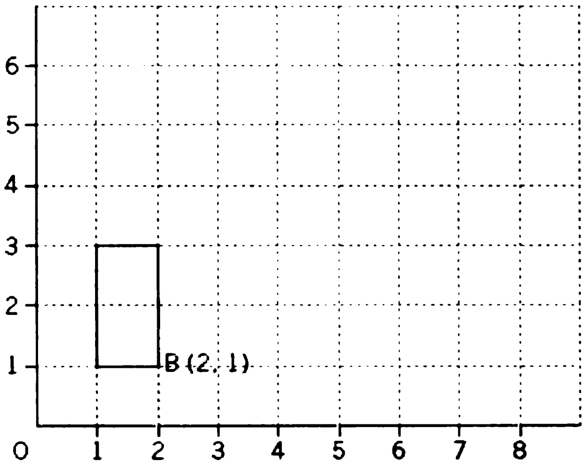
b Figure I and Figure III

c Figure II and Figure III

d none of the above

16. A rectangle is drawn on the first quadrant of a Cartesian grid as shown.

The coordinates of Point B are given in the diagram below.



If the rectangle is translated 4 units to the right and 3 units up, the new coordinates of Point B will be

a (3, 4)

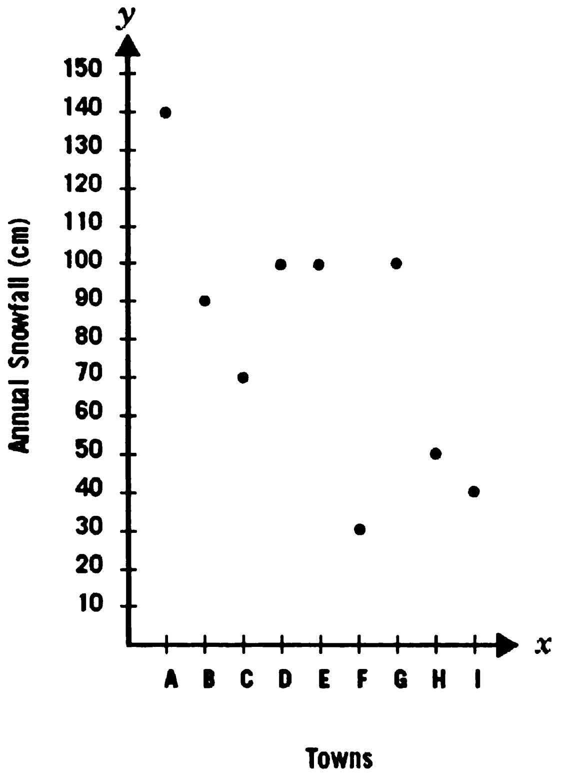
b (4, 3)

c (5, 5)

d (6, 4)

17. Nine towns, A to I, are spread out in a hilly region of the province. Their snowfall data for the past year are shown in a scatter plot.

What is the median annual snowfall for this group of towns?



a 80 cm

b 85 cm

c 90 cm

d 100 cm

18. One of the numbers below meets the following conditions:

- It is a composite number.
- It is between 22 and 32.
- It has more than 4 factors.
- It results in a prime number when its digits are added.

Which number is it?

a 23

b 25

c 28

d 30

19. If   = 4, what is the value of
21 – (4 ×   )?

a 5

b 13

c 16

d 68

20. Nicholas works on his math project for $2\frac{1}{2}$ hours on Monday, $1\frac{3}{4}$ hours on Wednesday and 3 hours on Thursday.

What is the total time he works on his math project, expressed in minutes?

a 235 minutes

b 360 minutes

c 361 minutes

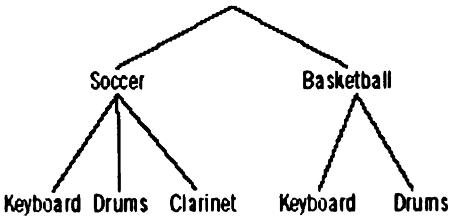
d 435 minutes

21. A school requires students to participate in a team sport and play a musical instrument.

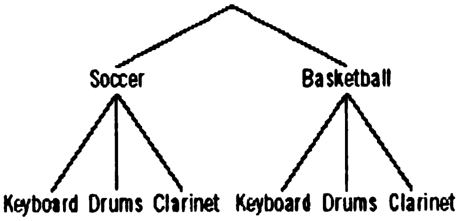
You are offered basketball and soccer as the team sports, and keyboard, drums and clarinet as the musical instruments.

Which diagram below shows all the possible choices for you?

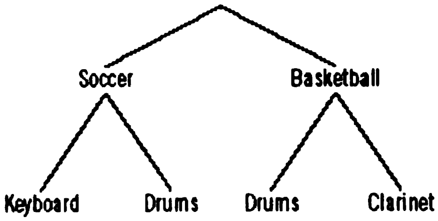
a



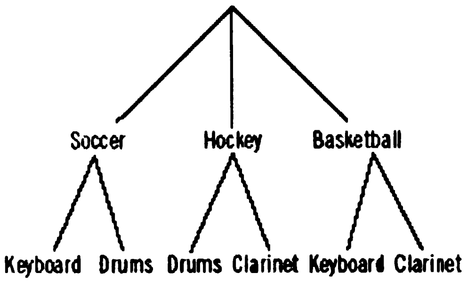
b



c



d



22. Kevin's batting average this baseball season is 0.346. Last season his batting average was 0.297. How much higher is Kevin's batting average this season?

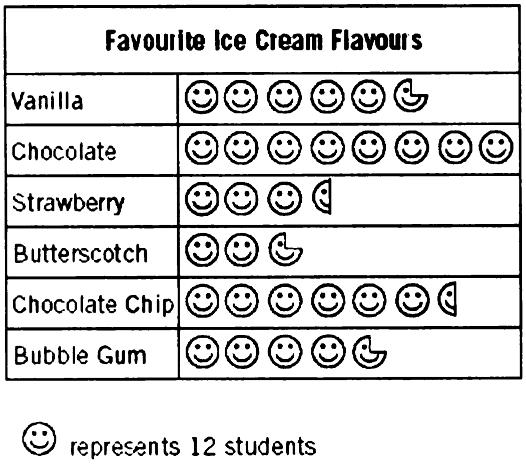
a 0.049

b 0.051

c 0.059

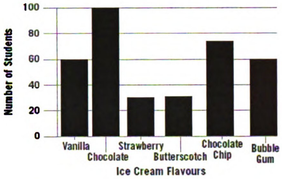
d 0.643

23. The pictograph below shows the number of students who chose different ice cream flavours as their favourite.



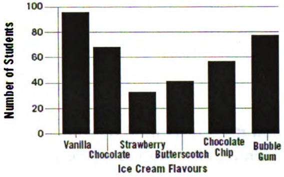
Which bar graph on the right best represents this data?

a



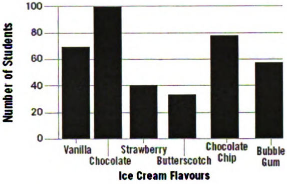
Number of Students
100
80
60
40
20
0
Vanilla
Chocolate
Strawberry
Butterscotch
Chocolate Chip
Bubble Gum
Ice Cream Flavours

b



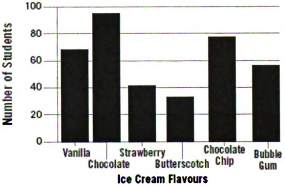
Number of Students
100
80
60
40
20
0
Vanilla
Chocolate
Strawberry
Butterscotch
Chocolate Chip
Bubble Gum
Ice Cream Flavours

c



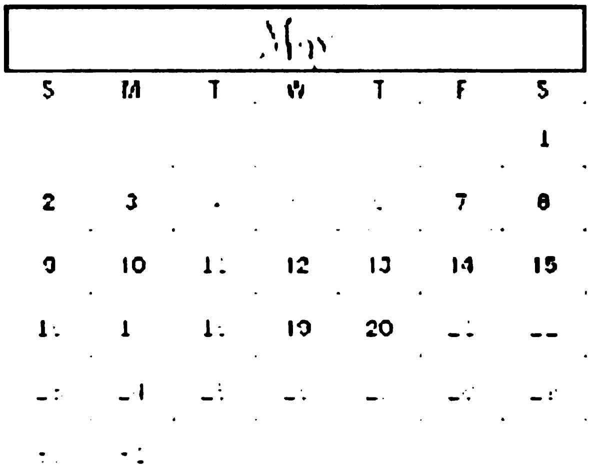
Number of Students
100
80
60
40
20
0
Vanilla
Chocolate
Strawberry
Butterscotch
Chocolate Chip
Bubble Gum
Ice Cream Flavours

d



Number of Students
100
80
60
40
20
0
Vanilla
Chocolate
Strawberry
Butterscotch
Chocolate Chip
Bubble Gum
Ice Cream Flavours

24. Which of the following is true of all three squares below?



May

S	M	T	W	T	F	S
						1
2	3	[illegible]	[illegible]	[illegible]	7	8
9	10	11	12	13	14	15
[illegible]	[illegible]	[illegible]	19	20	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]					

a Subtract 8 from the top left number to get the bottom right number.

b Add 8 to the top left number to get the bottom right number.

c Add 7 to the top left number to get the bottom right number.

d Subtract 7 from the top left number to get the bottom right number.

25. A watermelon has a mass of 2.4 kg. What is the mass expressed in g or mg?

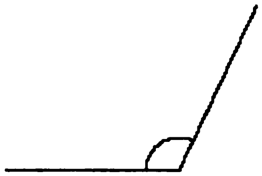
a 240 g or 240 000 mg

b 2400 g or 2 400 000 mg

c 2 400 000 g or 2400 mg

d 0.0024 g or 0.000 002 4 mg

26. The closest estimate of the angle shown is



a 70° to 80°.

b 85° to 95°.

c 110° to 120°.

d 140° to 150°.

27. Simaya's Grade 6 class has 25 students. The teacher tells her that this is 4% of the entire school's student population.

How many students are in her school?

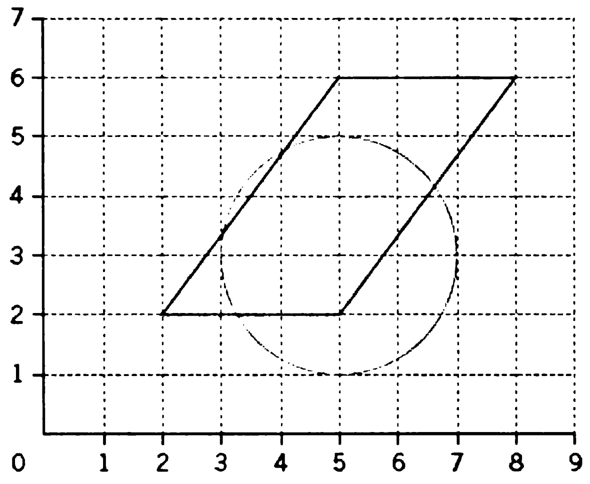
a 100

b 150

c 600

d 625

28. Which of the following descriptions is correct?



a The point (6, 2) is outside the circle and outside the parallelogram.

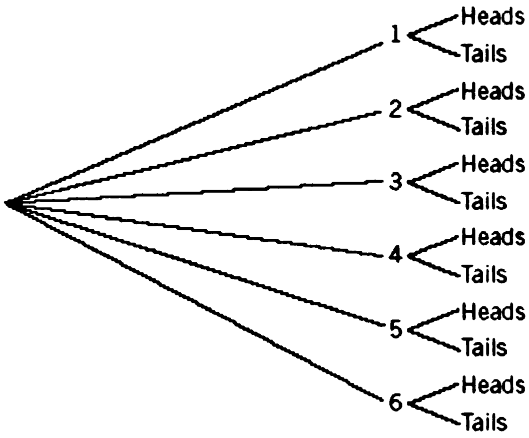
b The point (6, 2) is inside the circle and outside the parallelogram.

c The point (6, 2) is outside the circle and inside the parallelogram.

d The point (6, 2) is inside the circle and inside the parallelogram.

29. A number cube is tossed and a coin is flipped.

Using the tree diagram, determine which of the following events is most likely to occur.



a Tails appears on the coin.

b Heads appears on the coin, and a 4 appears on the number cube.

c A 3 appears on the number cube.

d Heads appears on the coin, and a 3 appears on the number cube.

30. A department store's sales for one week are listed below.

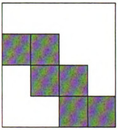
Clothing	$3240.00
Cosmetics	$900.00
Hardware	$2521.00
Appliances	$583.00
Other	$1011.00

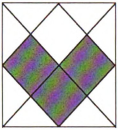
* Estimate the total sales for the week.

Explain your estimation strategy.

My estimate is _________.

31. Connor states that, for both diagrams, the shaded parts represent $\frac{3}{8}$ of the whole figure.

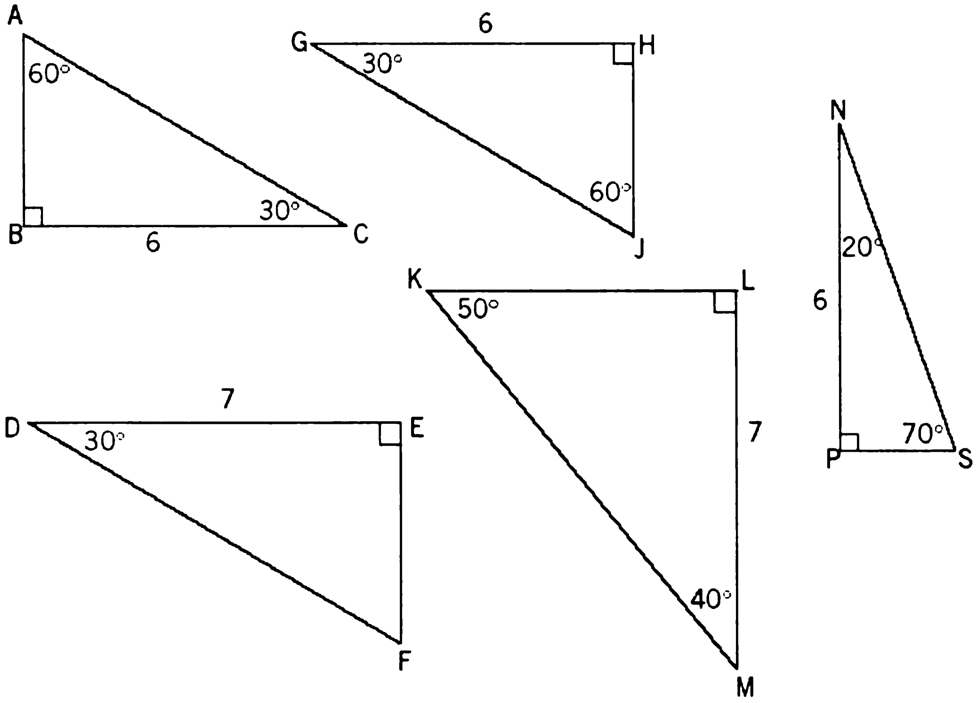




✱ Is Connor correct?

Justify your answer.

32. Lindsay is cutting triangles to use in making some paper crafts. She notices that some of the triangles are exactly the same.



✱ Find the congruent triangles.

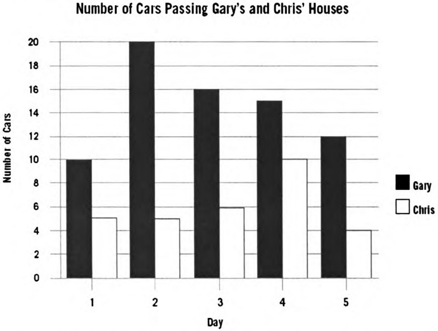
Justify your answer.

33. A. J. notices a pattern in the vertical posts and horizontal boards in his fence. He counts the number of vertical posts, then subtracts one and multiplies by two to find the number of horizontal boards.

- ✱ Fill in the table below to show the number of horizontal boards.

Number of Vertical Posts	Number of Horizontal Boards
2	
3	
4	
5	

34. Gary and Chris record the number of cars that pass their houses over 5 days. The results are shown below.



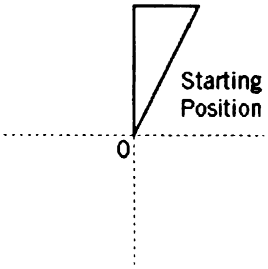
Gary says that each day the number of cars that pass his house is at least two times the number of cars that pass Chris' house.

Is Gary correct?

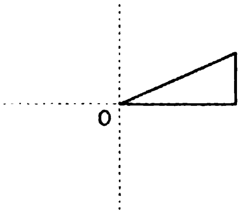
Explain your answer.

35. The triangle shown in the grid is rotated about point 0 by 90° in a clockwise direction.

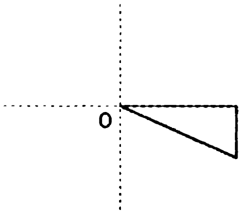
What is its new position?



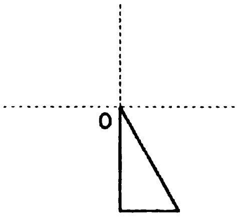
a



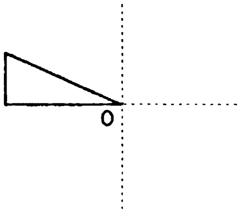
b



c



d



36. Andrew has $20 in nickels. He gives $3.75 to his sister and $1.15 to his grandma and he spends $0.65.

How many nickels does he have left?

a 237

b 289

c 14.45

d 0.7225

37. $\frac{1}{1} + \frac{22}{22} + \frac{333}{333} = ?$

a 1

b 3

c 6

d 356

38. Smita types 5400 words per hour.

How many words does she type per minute?

a 1.5 words per minute

b 90 words per minute

c 60 words per minute

d 5400 words per minute

39. Which number best completes the pattern below?

2, 5, 11, 23, 47, 95, ____

a 142

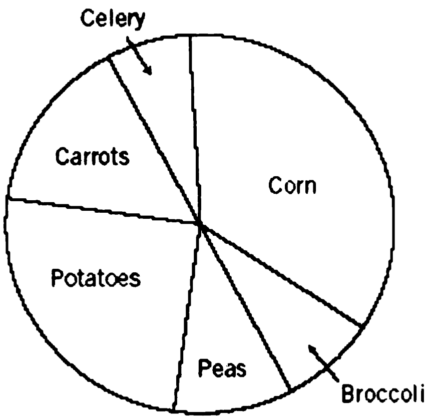
b 190

c 191

d 192

40. Mark does a survey on favourite vegetables in his class. The graph represents the response data.

Student Votes for Favourite Vegetables



Which vegetable was chosen by about 25% of the class?

a corn

b peas

c carrots

d potatoes

41. What is the least common multiple for 6 and 10?

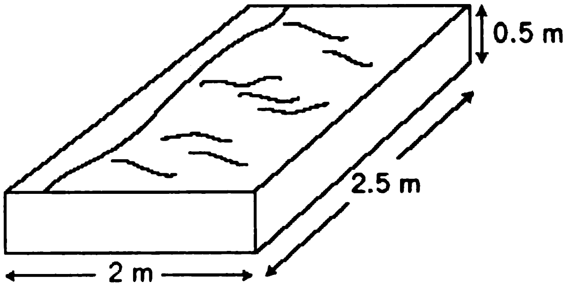
a 20

b 30

c 40

d 60

42. Alonzo's dad builds him a sandbox that measures 2.5 m long, 2 m wide and 0.5 m deep.



How much sand does Alonzo need to fill his sandbox to the top?

a 5 m^3

b 9 m^3

c 2.5 m^3

d 5.5 m^3

Appendix B

EQAO Grade 6 Mathematics Test 2006

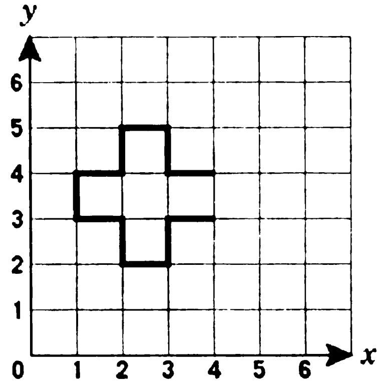
1 Which is the most appropriate unit of measurement to describe the area of the floor of a gym?

- a km^2
- b cm^3
- c m^2 *
- d m^3

2 Joseph has a measuring wheel that clicks once for every metre he walks. How many times will the wheel click when Joseph walks 2.6 km?

- a 2
- b 26
- c 260
- d 2600 *

3 Jacob draws most of an addition symbol on the Cartesian plane below.



Which two ordered pairs represent the location on the grid of the two points that should be connected to complete the addition symbol?

- **a** (3, 4) and (4, 4)
- **b** (4, 3) and (3, 3)
- **c** (3, 4) and (4, 3)
- **d** (4, 4) and (4, 3) *

4 Germaine buys one hamburger, one sandwich and two fruit salads.

Menu

Item	Amount
Hamburger	$3.50
Sandwich	$2.75
Fruit Salad	$1.60
Frozen Yogourt	$3.00

How much change should she receive from $20.00?

- **a** $9.15
- **b** $9.45
- **c** $10.55 *
- **d** $12.15

5 Which number, when placed in the box, makes the following number sentence true?

$15 - 6 \times 2 + 18 \div 3 = \square$

a 7

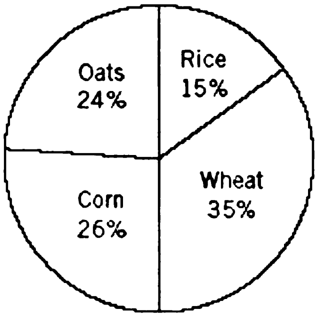
b 9 *

c 12

d 24

6 The graph below shows grain used to make cereal at a breakfast food factory.

Grain Used for Cereal



Based on the graph, which of the following statements is true?

a The amount of wheat used is more than the combined amount of corn and oats.

b The amount of corn used is more than the combined amount of oats and rice.

c The combined amount of wheat and rice used is the same as the combined amount of corn and oats. *

d The combined amount of oats and rice used is the same as the amount of wheat.

7 Examine the input-output table shown below.

Input	Output
2	5
3	8
4	11
6	17

Which of these rules describes the data?

a Multiply by 2 and add 1.

b Multiply by 4 and subtract 3.

c Multiply by 2 and add 5.

d Multiply by 3 and subtract 1.*

8 Pie is served at a picnic. Each pie is made up of 6 equal pieces. Bradley records the number of pieces each person eats in the table below.

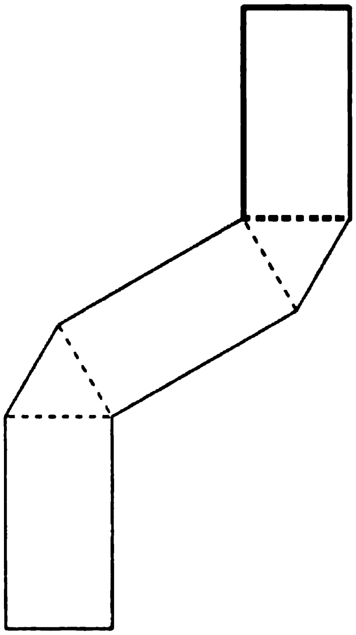
Name	Gurleen	Max	Ta-Shanya	Stewart	Brianne	Adrian
Number of Pieces Eaten	3	2	2	3	3	1

How many pies are eaten in total? Express your answer as a fraction.

Show your work.

They eat _______ pies.

9 Draw the three-dimensional figure that will be created when the following net is folded. Show all vertices and edges.



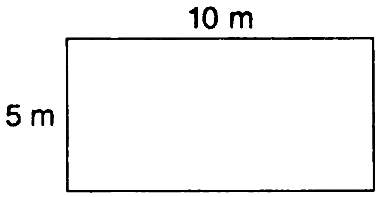
10 A spinner has 12 equal-sized sections. The sections are labelled 1 through 12.

What is the probability that Frieda will spin a multiple of 3 on her first spin?

Explain how you know.

The probability is ____________.

11 Susie wants to tile the floor of her family's rectangular play room. The tiles she plans to use are 10 cm by 10 cm squares. A drawing of the room is shown below.

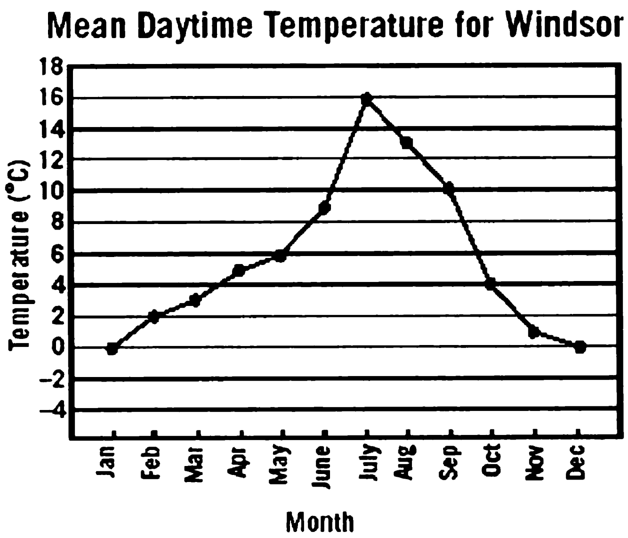


How many of the square tiles will Susie need to cover the floor of the play room?

Show your work.

Susie will need ___________ tiles.

12 The graph below shows the mean daytime temperature for Windsor.



Which month has a mean daytime temperature that is twice April's?

a July

b August

c September *

d October

13 To pick teams, the gym teacher puts the names of 8 boys and 6 girls in a bag, as shown below. The table shows the names.

Boys	Girls
Robert	Jessica
Ivan	Sarah
Hasan	Preija
Mohamed	Minon
Salvatore	Sunetra
Kieran	Ling
Paul	
Manuel	

The first 3 names picked at random from the bag were Paul, Jessica and Sarah. The names are not put back in. What is the probability that the next name picked at random will be a boy?

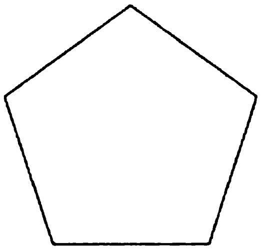
a $\frac{1}{2}$

b $\frac{7}{11}$ *

c $\frac{1}{7}$

d $\frac{8}{14}$

14 The regular pentagon shown below has 72° rotational symmetry.



How many 72° rotations will it take to return the vertices to their original positions?

- a 1
- b 2
- c 4
- d 5 *

15 A rectangular wall is being built. The table shows the dimensions of the wall after each day.

Wall Dimensions

Day	Height	Length
1	1 m	2 m
2	2 m	3 m
3	3 m	4 m
4	4 m	5 m

If the pattern continues, what will the perimeter of the wall be at the end of Day 10?

- a 42 m *
- b 38 m
- c 21 m
- d 19 m

16 The following pattern increases by following this rule: multiply the previous term by 3 and add 1.

5, 16, 49, 148, . . .

What is the next term in the sequence?

a 159

b 218

c 444

d 445 *

17 Which of the following is a factor of 70 but is not a prime number?

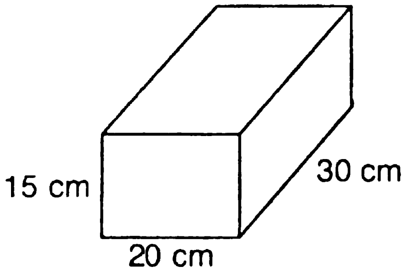
a 10 *

b 7

c 4

d 2

18 Four students calculate the volume of the shoe box shown below.



The following number sentences show the students' calculations. Which calculation is correct?

a 15 cm × 20 cm = 300 cm^2

b 20 cm × 30 cm = 600 cm^2

c 20 cm + 30 cm + 15 cm = 65 cm^3

d 15 cm × 20 cm × 30 cm = 9000 cm^3 *

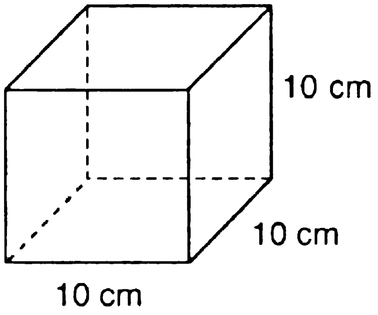
19 Which set is in order from least to greatest?

- **a** 1.153, 1.062, 0.13, 0.054
- **b** 0.13, 0.054, 1.162, 1.153
- **c** 0.054, 0.13, 1.153, 1.062
- **d** 0.054, 0.13, 1.062, 1.153 *

20 The results of a survey show that 30% of the people surveyed read a newspaper regularly. Which of the following numbers is equivalent to 30%?

- **a** 0.03
- **b** 3.0
- **c** $\frac{1}{3}$
- **d** $\frac{3}{10}$ *

21 A cube is shown below. It is 10 cm wide, 10 cm long and 10 cm high.



What is the area of one of the faces of the cube?

- **a** 10 cm^2
- **b** 30 cm^2
- **c** 100 cm^2 *
- **d** 1000 cm^2

22 Sam buys 4 items in a store. The mass of each item is recorded below.

9000 mg, 400 g, 0.04 kg, 0.009 kg

Which item has the greatest mass?

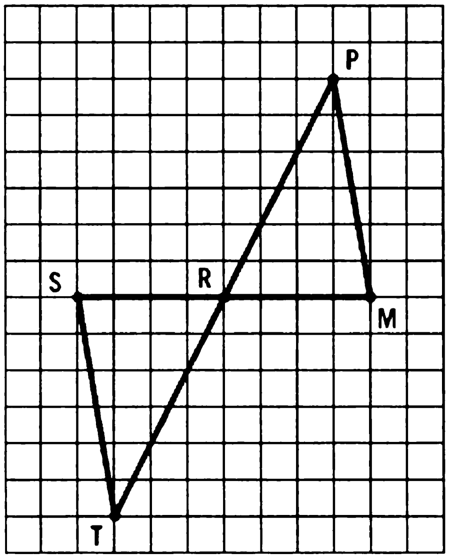
a 9000 mg

b 400 g *

c 0.04 kg

d 0.009 kg

23 Which answer best describes the transformation from ΔMPR to ΔRST?



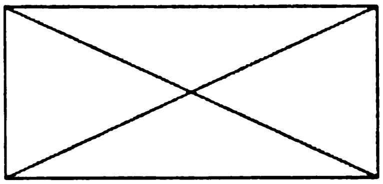
a Reflect about Point R.

b Rotate $\frac{1}{4}$ turn clockwise about Point M.

c Reflect about $\overline{RM}$.

d Rotate $\frac{1}{2}$ turn about Point R. *

24 A drawing of the back of an envelope is shown below.



Which statement best describes the back of the envelope?

a eight isosceles triangles

b four equilateral triangles

c a rectangle with two diagonals *

d a parallelogram surrounded by a rectangle

25 Cary needs to set up 144 chairs in rows. Each row must have an equal number of chairs. Which of the following could be the method Cary uses to set up the chairs?

a 14 rows of 10 chairs

b 12 rows of 14 chairs

c 6 rows of 21 chairs

d 8 rows of 18 chairs *

26 Johnna is planning a survey of students in her classroom. She wants to find their favourite food for lunch at school. Which of the following would be the best question for Johnna to ask in her survey?

a "What is your favourite food?"

b "What are your friends' favourite foods?"

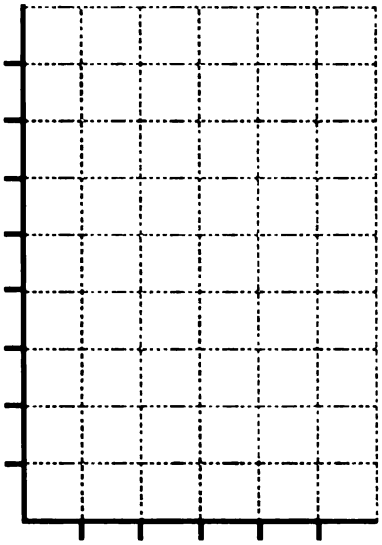
c "What is your favourite food for lunch at school?" *

d "What is your favourite food—a sandwich or soup?"

27 Ranjit makes the chart below to record the amount of money collected during a fundraising event.

Day	Monday	Tuesday	Wednesday	Thursday	Friday
Amount of Money Collected	$50	$125	$75	$25	$175

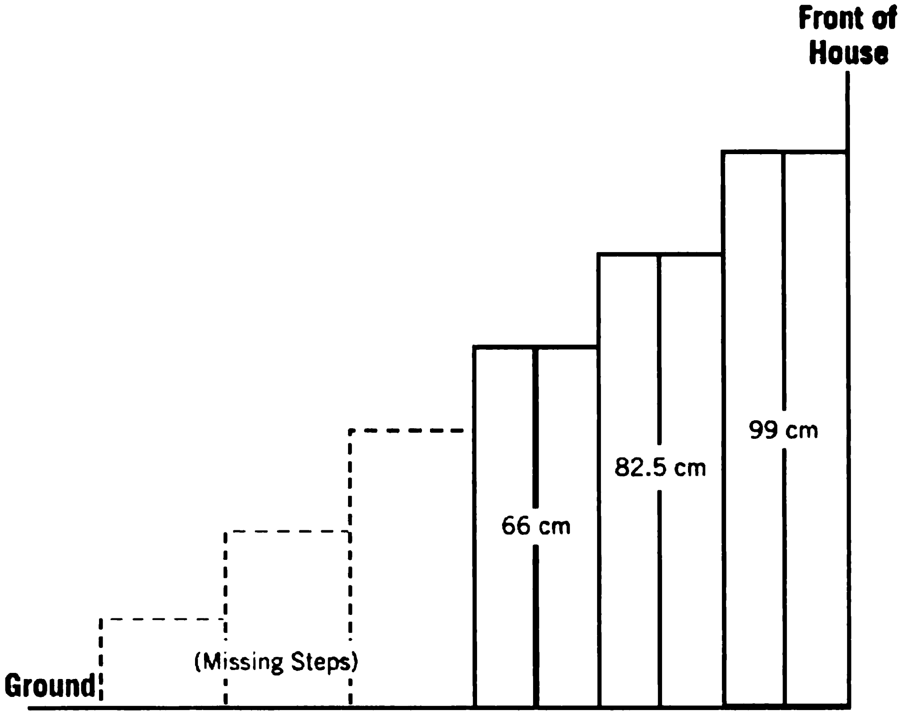
Make a broken-line graph to represent the data. Remember to include all titles and labels.



Explain your choice of scale.

28 A carpenter is replacing some missing steps at the front of Dena's house. The bottom three steps are missing. He wants to use the same heights for the new steps as the old steps. The carpenter measures the height from the ground to the top of each remaining step.

- The fourth step is 66 cm from the ground.
- The fifth step is 82.5 cm from the ground.
- The sixth step is 99 cm from the ground.



The carpenter plans to make each step increase by the same amount.

What are the heights of the first, second and third steps?

Show or explain your work.

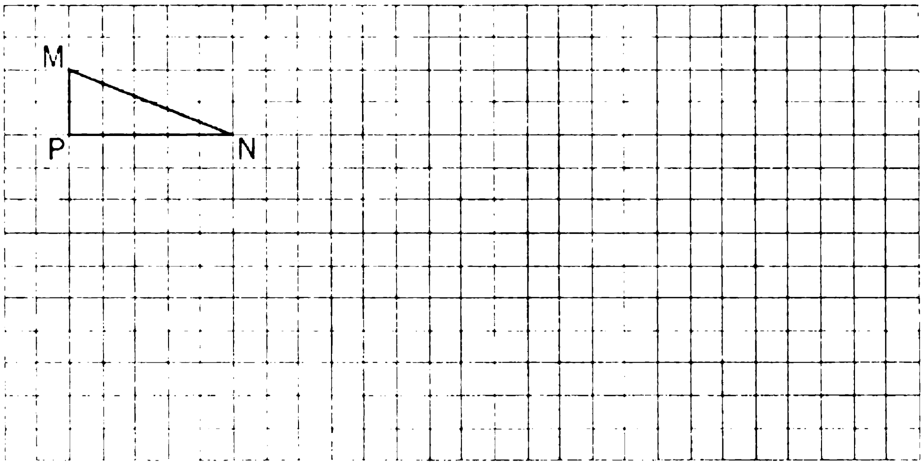
28 The rectangular ceiling of a room has an area of 36 m^2. The ceiling needs 3 coats of paint. Each can of paint covers 25 m^2.

About how many cans of paint are needed to paint the ceiling?

Explain your thinking.

_____ cans of paint are needed.

30 Use two transformations of different types to move the triangle on the grid below to a new position. Show both transformations and label M, N and P on the new figure.



Explain your two transformations, using the correct name for each transformation.

31 In a hockey arena, the first row has 276 seats, the second row has 288 seats and the third row has 300 seats. Each row after this continues to increase by the same number. If the arena has a total of 6 rows, how many seats are in the arena?

a 1836 *

b 1176

c 972

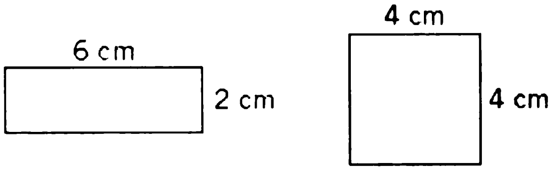
d 312

32 Ms. Vanstone asks her students to draw a rectangle and a square with the areas and perimeters given below.

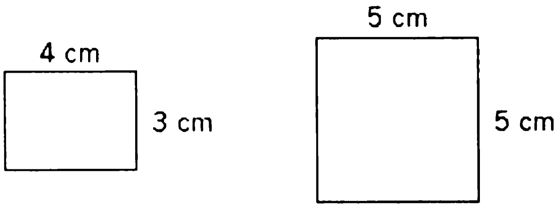
	Rectangle	**Square**
Area	12 cm^2	25 cm^2
Perimeter	16 cm	20 cm

Which shows two correct drawings?

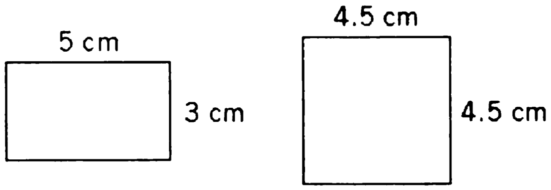
a



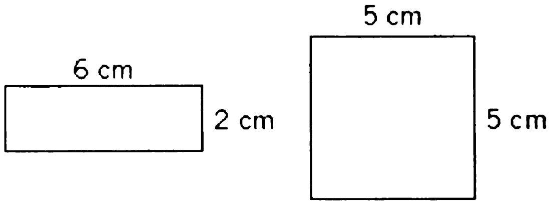
b



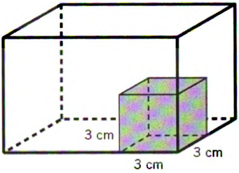
c



d *



33 Twelve cubes measuring 3 cm by 3 cm by 3 cm fit perfectly into the rectangular prism shown below.



What is the volume of the rectangular prism in cm^3?

a 36 cm^3

b 162 cm^3

c 288 cm^3

d 324 cm^3 *

34 What value, when placed in the box, would make the following equation true?

6 × □ − 4 = 56 + 6

a 10

b 11 *

c 31

d 62

35 The same number is added to each term in a pattern to get the value of the next term. Below are the fourth, fifth and sixth terms in the pattern.

... 95, 98, 101, ...

What are the first, second and third terms in the pattern?

- **a** 83, 85, 87
- **b** 83, 86, 89
- **c** 86, 88, 92
- **d** 86, 89, 92 *

36 Chloe's parents are buying a car. They want to pick 1 colour at random from 4 possible car colours. Which of the following methods should they use?

- **a** Flip a coin.
- **b** Toss a 6-sided number cube with 1 through 6 on the faces.
- **c** Use a spinner with 4 equal-sized sections labelled with the 4 possible colours. *
- **d** Pick one card from 10 cards with 1 of the 4 colours written on each face.

Reference

Abedi, J. (2004). The No Child Left Behind Act and English language learners: assessment and accountability issues. *Educational Researcher*, 33, 4-14.

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. (1994). Using multidimensional item response theory to understand What items and tests are measuring. *Applied Measurement in Education,* 7, 255-278.

Ackerman, T.A., Gierl, M. G., and Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, 22, 37-51.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 62, 317-332.

Baker, F.B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement*, 16, 87-96.

Baker, F.B. (1993). Equating tests under the nominal response model. *Applied Psychological Measurement*, 17, 239-251.

Briggs, D.C. and Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Educational Measurement,* 4 , 87-100.

Carlson, J. E. (1987). Multidimensional item response theory estimation: A computer program (ACT Research Report. No. 87-190). Iowa City, IA: American College Testing Program.

Cohen, A.S., & Kim, S.-H. (1998). An investigation of linking methods under the graded response model. *Applied Psychological Measurement*, 22, 116-130.

Darling-Hammond, L. (2002). Standards, assessment, and educational policy: in pursuit of genuine accountability. The eighth annual William H. Angoff Memorial Lecture, Princeton, NJ: Educational Testing Services.

Donoghue J.R. (1993) An empirical examination of the IRT information in polytomously scored reading items. ETS research report. RR-93-12. Princeton, NJ: Educational Testing Services.

Edwards, M. C. and Vevea, J. L. (2006). An empirical Bayes approach to subscore augmentation: how much strength can we borrow? *Journal of Educational and Behavioral Statistics,* 31, 241-259

Gower, J.C. and Dijksterhuis, G.B.(2004). *Procrustes Problems.* Oxford University Press: Oxford:

Haberman, S. J. (2005). Interpretation of reliability. ETS research report, RR-05-29. NJ: Educational Testing Services.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22,* 144-149.

Hambleton, R.K. (1989) Principles and selected application of item response theory. *Educational Measurement.* 3rd. ed. New York: American Council on Education & Macmillan Pub. Co.

Hattori, T. (1998). Equating the parameters for the generalized partial credit model-minimum and item characteristic curve methods. Proceedings of the 62nd annual meeting of the Japanese Psychological Association, p. 417. (in Japanese)

Herman, J.L., Webb, N.M., & Zuniga, S.A. (2005). Measurement issues in the alignment of standards and assessments: a case study. CSE Report 653. Center for the Study of Evaluation National Center for Research on Evaluation, Standards, and Student Testing. Retrieved on Aug. 2007 from http://www.cse.ucla.edu/products/reports/r653.pdf

Hill, R. K. & DePascale, C.A. (2002) Reliability of No Child Left Behind Accountability Designs. *Educational Measurement: Issues and Practice,* 22 (3), 12–20.

Hirsch, T.M. (1988). Multidimensional equating. Unpublished doctoral dissertation. Florida State University.

Hirsch, T.M. (1988). Multidimensional equating. *Journal of Educational Measurement,* 26, 337-349.

Holland, P.W. and Dorans, N. J. (2006). Linking and equating. In Brennan, R.L. (Eds.) *Educational Measurement.*4th ed. (pp. 187-220). West Post, CT: American Council on Education & Praeger Publisher.

Horton, M.T. and Hanes, S.M. A user's guide to preparing submissions for the NCLB standards and assessments peer review. Prepared for the Office of Elementary and Secondary Education, U.S. Department of Education. Retrieved on Dec. 2006 from http://www.ed.gov/admins/lead/account/peerreview/usersguide.doc.

Jackson, M. (2007). Large-scale assessment: supporting the everyday work of schools. Retrieved on February, 2007 from http://www.eqao.com/Publications/ArticleReader.aspx?Lang=E&article=b07A001&section

Jager, R. (1991). A comparison of compensatory, conjunctive, and disjunctive models for weighing attributes of school quality. Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991). (ERIC Document Reproduction Service No. ED348732)

Jodoin, M. G., Keller L., and Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *The Journal of Experimental Education*, 71(3), 229-250.

Kelderman, H. and Rijkes, C.P.M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika*, 59:149-176.

Kim, J-S. and Hanson, B. A. (2002) Test equating under the multiple-choice model. *Applied Psychological Measurement,* 26 (3), 255–270.

Kim, J-P. (2001) "Proximity measures and cluster analysis in multidimensional item response theory", Unpublished doctoral dissertation. Michigan State University.

Kim, S. (2004). Unidimensional IRT scale linking procedures for mixed-format test and their robustness to multidimensionality. Unpublished doctoral dissertation. University of Iowa.

Kim, S. and Kolen, M.J. (2004a). STUIRT: A Computer Program for Scale Transformation under Unidimensional Item Response Theory Models. Version 1.0. Downloadable at http://www.education.uiowa.edu/casma/IRTPrograms.htm

Kim, S. and Kolen, M.J. (2004b). A Manual of the STUIRT program. Downloadable at http://www.education.uiowa.edu/casma/IRTPrograms.htm

Kim, S. and Lee, W. (2004). IRT scale linking methods for mixed-format tests (ACT Research paper) Iowa City, IA: ACT, inc.

Kolen, M.l J. and Brennan, R.L. (2004) *Test Equating, Scaling, and Linking: Methods and Practices.* New York: Springer, c2004. 2nd ed.

Lane, S. (2005). Status and future directions for performance assessments in education. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Li, Tianli (2006). The Effect of Dimensionality on Vertical Scaling. Unpublished doctoral dissertation, Michigan State University.

Li, Y.H. and Lissitz, R.W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement*, 24, 115-138.

Li, Y.H., Lissitz, R.W., and Yang, Y. (1999). Estimating IRT equating coefficients for tests with polytomously and dichotomously scored items. Paper presented at the annual meetings of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Linn, R.L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 6, 83-102.

Linn, R.L. (2003). Accountability: responsibility and reasonable expectations. CSE Report 601. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Linn, R.L. (2005) Fixing the NCLB Accountability System. CRESST Policy Brief 8. Summer 2005. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

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

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*, 169-194.

Lukhele, R., Thissen, D., and Wainer, H., (1993) On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. ETS report, RR-93-06. Princeton, NJ: Educational Testing Services.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*, 139-160.

Martin,S. E., del Pino, G., De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement* 30, 183-203.

Martineau, J. A., Mapuranga, R., & Ward, K. (2006). Confirming content structure in standardized state assessment using multidimensional item response theory. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA. Apr. 7-11, 2006.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.

Masters, G. N. and Wright, B.D. (1997). The partial credit model. In W.J. van der Linden and R.K. Hambleton (Eds.), Handbook of modern item response theory (pp. 101-121). New York: Springer.

Mehrens, W.A. and Phillips, S.E. (1989). Using College GPA and Test Scores in Teacher Licensure Decisions: Conjunctive versus Compensatory Models. *Applied Measurement in Education*, 2 (4), 277-88

Miller, T.R. and Hirsch, T.M. (1992) Cluster analysis of angular data in applications of multidimensional item response theory. *Applied Measurement in Education* 5, 193-211

Min, K. (2003). The impact of scale dilation on the theory of linking of multidimensional item response theory calibrations. Unpublished doctoral dissertation. Michigan State University.

Mislevy, R.J. (1992). *Linking Educational Measurements: Concepts, Issues, Methods, and Prospects.* Princeton, NJ: Educational Testing Service, Policy Information Center.

Monaghan, W. (2007). The facts about subscores. *ETS R&D Connections* 4, July 2006. Retrieved on Aug. 2007 from http://www.ets.org/Media/Research/pdf/RD_Connections4.pdf

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm, *Applied Psychological Measurement,*16 (2), 159-176.

Muraki, E. (1997). A generalized partial credit model. In van der Linden, W. J. & Hambleton, R.K. (1997). *Handbook of Modern Item Response Theory*. (pp. 153-164). New York, N.Y.: Springer.

Muraki, E., and Bock, R.D. (2002) PARSCALE 4: IRT based item analysis and test scoring for rating-scale data. Chicago, IL: Scientific Software International, Inc.

National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics*. Retrieved in 2003, from http://standards.nctm.org/document/index.htm

Oshima, T.C., Davey, T.C., and Lee, K. (2000). Multidimensional linking: four practical approaches. *Journal of Educational Measurement*, 37, 357-373

Paulsen, C.A., Ferrara, S., Birns, J. and Leclerc, K. J. (2002) Multiple measures for student assessment and accountability in Massachusetts. A paper prepared in support of Contract #SC DOE 1100 2ERRC71 Massachusetts Education Reform Review Commission. Concord, MT: American Institutes for Research.

Reckase, M.D. (1985) The difficulty of test items that measure more than one ability. *Applied Psychological Measurement,* 9, 401-412.

Reckase, M.D. and McKinley, R.L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement,* 15, 361-373

Reckase, M.D. and Martineau, J. (2004). The vertical scaling of science achievement tests. Paper commissioned by the Committee on test design for K_12 science achievement, Center for Education, National Research Council

Reckase MD (2005). Multidimensional item response theory models. In Kimberly Kempf-Leonard (Ed.) *Encyclopedia of Social Measurement,* (Vol. 2, pp. 771-777). San Diego, Calif.; London: Academic.

Reckase, M.D. (2006). Dimensions, coordinates, and hypothetical constructs from MIRT calibrations. Paper presented at the annual meeting of the Psychometric Society, Montreal, Canada.

Reckase, M.D. (in press) *Multidimensional Item Response Theory.*

Samejima, F. (1969). Estimation of a latent ability using a response pattern of graded scores. *Pychometric Monograph Supplement*, 34 (Monograph No. 17).

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

Sinharay, S., Haberman, S. & Puhan, G. (2007). Subscores Based on Classical Test Theory: To Report or Not To Report. *Educational Measurement: Issues and Practice*, 26 (4), 21–28.

Stake, R. (1995). The invalidity of standardized testing for measuring mathematics assessment. In Thomas A. Romberg (Ed.) *Reform in School Mathematics and Authentic Assessment*, State University of New York Press: Albany, NY.

Stocking, M. L., and Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.

Taherbhai, H.M. and Seo, D.Y. (2007). Comparing concurrent versus fixed parameter equating with common items: using the dichotomous and partial credit models in a mixed-item format test. *Journal of Applied Measurement*, 8, 84-96.

Tate, R.L. (2000) Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement*, 37, 329-346.

Thissen, D. and Edwards, M.C. (2005). "Diagnostic scores segmented using multidimensional item response theory: preliminary investigation of MCMC strategies. A paper presented at the annual meeting of the National Council on Measurement in Education in Montreal, PQ, Canada, April 12-14, 2005.

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

Volante, L. (2004). Teaching to the test: what every educators and policy-makers should know. *Canadian Journal of Educational Administration and Policy*, Issue 35, September 25, 2004.

Volante, L. (2006). Standards-based reform: can we do better? *Education Canada*, 47 (1), 54-56. Winter 2006/2007

Wang, M. (1985) Fitting a unidimensional model to multidimensional item response data: the effects of latent space misspecification on the application of IRT. ACT Research Paper 6-24-85. Iowa City, IA: American College Testing Program.

Wang, M. (1986) Fitting a unidimensional model to multidimensional item response data. Paper presented at the Office of Naval Research Contractors Meeting, Gatlinburg, TN.

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

Yao, L. (2003). BMIRT: Bayesian multivariate item response theory [Computer software]. Monterey, CA: CTB/McGraw-Hill.

Yao, L. and Schwarz, R.D. (2006). A multidimensional partial credit model with associated item and test statistics: an application to mixed-format tests. *Applied Psychological Measurement*, 30(6), 369-492.

Yao, L. and Boughton, K.A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, 31(2), 83-105.

Yao, L. and Boughton, K.A. (2006) Approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, 30, 469-497.

Yon, Haniza. (2006). Multidimensional Item Response Theory (MIRT) Approaches to Vertical Scaling. Unpublished doctoral dissertation. Michigan State University.

Thissen, D. and Edwards, M.C. (2005). "Diagnostic scores segmented using multidimensional item response theory: preliminary investigation of MCMC strategies. A paper presented at the annual meeting of the National Council on Measurement in Education in Montreal, PQ, Canada, April 12-14, 2005.

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

Volante, L. (2004). Teaching to the test: what every educators and policy-makers should know. *Canadian Journal of Educational Administration and Policy*, Issue 35, September 25, 2004.

Volante, L. (2006). Standards-based reform: can we do better? *Education Canada*, 47 (1), 54-56. Winter 2006/2007

Wang, M. (1985) Fitting a unidimensional model to multidimensional item response data: the effects of latent space misspecification on the application of IRT. ACT Research Paper 6-24-85. Iowa City, IA: American College Testing Program.

Wang, M. (1986) Fitting a unidimensional model to multidimensional item response data. Paper presented at the Office of Naval Research Contractors Meeting, Gatlinburg, TN.

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

Yao, L. (2003). BMIRT: Bayesian multivariate item response theory [Computer software]. Monterey, CA: CTB/McGraw-Hill.

Yao, L. and Schwarz, R.D. (2006). A multidimensional partial credit model with associated item and test statistics: an application to mixed-format tests. *Applied Psychological Measurement*, 30(6), 369-492.

Yao, L. and Boughton, K.A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, 31(2), 83-105.

Yao, L. and Boughton, K.A. (2006) Approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, 30, 469-497.

Yon, Haniza. (2006). Multidimensional Item Response Theory (MIRT) Approaches to Vertical Scaling. Unpublished doctoral dissertation. Michigan State University.